Statistical Methodology 7 (2010) 240–253

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors

Adrian Dobra [a,b,*], Héléne Massam [c]

[a] Department of Statistics, University of Washington, Seattle, WA 98155-4322, USA

[b] Department of Biobehavioral Nursing and Health Systems, School of Nursing, University of Washington, Seattle, WA 98155-4322, USA

[c] Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON, M3J 1P3, Canada

ARTICLE INFO

Article history:
Received 26 October 2008
Received in revised form
10 April 2009
Accepted 10 April 2009

Keywords:
Bayesian analysis
Contingency table
Hierarchical log-linear model
Markov chain Monte Carlo
Model selection
Stochastic search

ABSTRACT

We describe a novel stochastic search algorithm for rapidly identifying regions of high posterior probability in the space of decomposable, graphical and hierarchical log-linear models. Our approach is based on the Diaconis–Ylvisaker conjugate prior for log-linear parameters. We discuss the computation of Bayes factors through Laplace approximations and the Bayesian iterative proportional fitting algorithm for sampling model parameters. We use our model determination approach in a sparse eight-way contingency table.

1. Introduction

Many datasets arising from social studies, clinical trials or, more recently, genome-wide association studies can be represented as multi-way contingency tables. Log-linear models [1] are a common way to summarize the most relevant interactions that exist among the variables involved. Determining those log-linear models that are best supported by the data is a problem that has been studied in the literature [2–4]. When the number of observed samples is considerable with respect to the number of cells in the table, asymptotic approximations to the null distribution of the generalized likelihood ratio test statistic lead to appropriate results. However, in the case of sparse contingency tables that contain mostly counts of zero, the large sample assumptions no longer hold, hence using the same types of tests might lead to unsuitable results. The number of degrees of freedom associated with a

* Corresponding author at: Department of Statistics, University of Washington, Seattle, WA 98155-4322, USA.
E-mail addresses: adobra@u.washington.edu (A. Dobra), massamh@yorku.ca (H. Massam).

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2009.04.002

log-linear model has to be properly adjusted as a function of the zero counts, while some log-linear parameters become non-identifiable due to the non-existence of the maximum likelihood estimates — see [5] for a discussion.

The Bayesian paradigm to model selection avoids these issues through the specification of prior distributions for model parameters [6]. Markov chain Monte Carlo (MCMC) algorithms have been traditionally used to identify models with high posterior probability. Dellaportas and Forster [7] is a key reference that describes a reversible jump MCMC method applied to decomposable, graphical and hierarchical log-linear models. Other notable papers that develop various MCMC schemes for discrete data include [8–12].

While MCMC methods seem to work well for problems involving a relatively small number of candidate models, they tend to be less efficient as the dimensionality of the model space grows exponentially. Jones et al. [13] and Hans et al. [14] highlight this issue in the context of Gaussian graphical models and regression variable selection. They introduce the shotgun stochastic search (SSS) method that is similar to MCMC but it focuses on aggressively moving towards regions of high posterior probability in the model space instead of attempting to sample from the posterior distribution over the model space.

The aim of this paper is to present a novel stochastic search method for decomposable, graphical and hierarchical log-linear models which we call the mode oriented stochastic search (MOSS). The essence of MOSS is the identification of models such that the ratio of their posterior probability and the posterior probability of the best model is above a certain threshold. This is the set of models in Occam's window, as defined by Madigan and Raftery [8]. MOSS requires an efficient computation of the marginal likelihood of models in the search space. Such a computation is made possible through the use of the Diaconis–Ylvisaker conjugate prior for "baseline" log-linear parameters for hierarchical models. This conjugate prior has been studied in detail in [15]. Using this conjugate prior is indeed crucial because it allows us to produce the mode of the high-dimensional joint posterior distribution of log-linear parameters using the iterative proportional fitting (IPF) algorithm. This in turn allows us to compute the Laplace approximation to the marginal likelihood of hierarchical log-linear models. Another advantage of this conjugate prior is that, as we show in Section 3, it is the conjugate prior to an exponential family, hence sampling from the posterior distribution of the log-linear parameters can be done using the Bayesian iterative proportional fitting algorithm originally proposed by Piccioni [16].

The structure of the paper is as follows. In Section 2 we introduce the example used throughout the paper. In Section 3 we give a Diaconis–Ylvisaker conjugate prior distribution for the log-linear parameters together with some of its main features, while in Section 4 we show how to compute the marginal likelihood of decomposable, graphical and hierarchical models based on this prior. In Section 5 we present our new stochastic search method, discuss its properties and apply it to two examples. In Section 6 we give the details of the Bayesian iterative proportional fitting algorithm for polychotomous variables. In Section 7 we give some concluding comments.

2. Motivating example: Household study in Rochdale

We consider a cross-classification of eight binary variables relating women's economic activity and husband's unemployment from a survey of households in Rochdale. This study was conducted to elicit information about factors affecting the pattern of economic life and their time dynamics — see [4] page 279. The variables are as follows: a, wife economically active (no, yes); b, age of wife >38 (no, yes); c, husband unemployed (no, yes); d, child ≤ 4 (no, yes); e, wife's education, high-school+ (no, yes); f, husband's education, high-school+ (no, yes); g, Asian origin (no, yes); h, other household member working (no, yes). There are 665 individuals cross-classified into 256 cells, which means that the resulting table is sparse having 165 counts of zero, 217 counts with at most three observations, but also a few large counts with 30 or more observations.

3. Conjugate priors for hierarchical log-linear models

In the Bayesian model selection framework, the choice of a prior distribution is made on the basis of, first, availability and ability to reflect prior knowledge and, next, mathematical convenience

whenever possible. If the search is restricted to the class of discrete models Markov with respect to an undirected decomposable graph G, it is convenient to use the hyper-Dirichlet distribution as defined by Dawid and Lauritzen [17]. The hyper-Dirichlet distribution is a conjugate prior for the clique and separator marginal cell counts of the multinomial distribution Markov with respect to G. Its hyper-parameters can be thought of as representing the clique and separator marginal cell counts of a fictive prior table of counts and they give enough flexibility for the representation of prior beliefs — for example, see [8] or [9].

When the class of possible models considered is the more general class of graphical Markov models with respect to any undirected graph, or the even wider class of hierarchical models, the only priors available in the literature so far are normal priors for the log-linear parameters. Knuiman and Speed [18] use a multivariate normal prior for the log-linear parameters. Dellaportas and Forster [7] use a variant of this prior. King and Brooks [19] propose another multivariate normal prior for the log-linear parameters which has the advantage that the corresponding prior distribution on the cell counts can also be derived explicitly. Recently Massam et al. [15] have expressed the multinomial distribution in terms of random variables which are all possible marginal counts rather than the cell counts. They also developed and studied the corresponding conjugate prior as defined by Diaconis and Ylvisaker [20] (henceforth abbreviated the DY conjugate prior) for the log-linear parameters for the general class of hierarchical log-linear models.

In this section we show how to derive the DY conjugate prior for log-linear parameters and some of its main properties.

3.1. Model parametrization

Let V be the set of criteria defining the contingency table. Denote the power set of V by $\mathcal{E}$ and take $\mathcal{E}_{\ominus} = \mathcal{E} \setminus \{\emptyset\}$. Let $X = (X_\gamma, |\ \gamma \in V)$ such that X_γ takes its values (or levels) in the finite set I_γ of dimension $|I_\gamma|$. When a fixed number of individuals are classified according to the $|V|$ criteria, the data are collected in a contingency table (n) with cells indexed by combination of levels for the $|V|$ variables. We adopt the notation of [21] and denote a cell by $i = (i_\gamma, \gamma \in V) \in \mathcal{I} = \times_{\gamma\in V} \mathcal{I}_\gamma$. The count in cell i is denoted $n(i)$ and the probability of an individual falling in cell i is denoted $p(i)$. We write $(n) = (n(i), i \in \mathcal{I})$ and $(p) = (p(i), i \in \mathcal{I})$. The grand total of (n) is $N = \sum_{i\in\mathcal{I}} n(i)$, while the grand total of (p) is 1. For $E \subset V$, cells in the E-marginal table (n_E) are denoted $i_E \in \mathcal{I}_E = \times_{\gamma\in E} \mathcal{I}_\gamma$. The marginal counts in (n_E) are denoted $n(i_E)$, $i_E \in \mathcal{I}_E$. The counts (n) follow a multinomial Mult$(N; (p))$ distribution with density function proportional to

$$g((n), (p)) = \prod_{i\in\mathcal{I}} p(i)^{n(i)}. \tag{1}$$

Let i^* be a fixed but arbitrary cell that we take to be the cell indexed by the "lowest levels" of each factor. We denote these lowest levels by 0. Therefore i^* can be thought of as the cell $i^* = (0, 0, \ldots, 0)$. We define the log-linear parameters to be

$$\theta(i_E) = \sum_{F\subseteq E} (-1)^{|E\setminus F|} \log p(i_F, i^*_{F^c}), \tag{2}$$

which, by the Moebius inversion formula, is equivalent to

$$p(i_E, i^*_{E^c}) = \exp \sum_{F\subseteq E} \theta(i_F). \tag{3}$$

We denote $\theta(i^*) = \theta(i_\emptyset) = \theta_\emptyset$ and $p(i^*) = p_\emptyset$. Remark that $p_\emptyset = \exp\theta_\emptyset$. It is easy to see that the following lemma holds.

Lemma 3.1. *If for $\gamma \in E$, $E \subseteq V$ we have $i_\gamma = i^*_\gamma = 0$, then $\theta(i_E) = 0$.*

This result shows that our parametrization is the "baseline" or "corner" constraint parametrization that sets to zero the values of the E-interaction log-linear parameters when at least one index in E is at level 0 — see [3]. Therefore, for each $E \subseteq V$, there are only $d_{\mathcal{D}} = \prod_{\gamma\in E}(|\mathcal{I}_\gamma| - 1)$ parameters and

for any $E \subseteq V$, we define $\mathcal{I}_E^* = \{i_E \mid i_\gamma \neq i_\gamma^*, \forall \gamma \in E\}$. We denote $\mathcal{I}^* = \mathcal{I} \setminus \{i^*\}$. We use the notation $F \subseteq_\ominus E$ to express that F is included in E but is not equal to the empty set and, for $i_E \in \mathcal{I}_E^*, E \in \mathcal{E}$, we write $i(E) = (i_E, i_{E^c}^*)$. The notation $i(E)$ refers to the cell having components $i_\gamma \neq 0, \gamma \in E$ and $i_\gamma = 0, \gamma \in E^c$ and should not be confused with the cell i_E in the E-marginal table.

From (3) we obtain the following expression for the cell probabilities in terms of the log-linear parameters

$$p_\emptyset = \frac{1}{1 + \sum_{E \in \mathcal{E}_\ominus} \sum_{i_E \in \mathcal{I}_E^*} \exp\left(\sum_{F \subseteq_\ominus E} \theta(i_F)\right)}, \tag{4}$$

and

$$p(i(E)) = \frac{\exp \sum_{F \subseteq_\ominus E} \theta(i_F)}{1 + \sum_{E \in \mathcal{E}_\ominus} \sum_{i_E \in \mathcal{I}_E^*} \exp\left(\sum_{F \subseteq_\ominus E} \theta(i_F)\right)}, \quad E \in \mathcal{E}_\ominus. \tag{5}$$

3.2. The multinomial for hierarchical log-linear models

Consider the hierarchical log-linear model m generated by the class $\mathcal{A} = \{A_1, \ldots, A_k\}$ of subsets of V which, without loss of generality, can be assumed to be maximal with respect to inclusion. We write $\mathcal{D} = \{E \subseteq_\ominus A_i, i = 1, \ldots, k\}$ for the indexing set of all possible interactions in the model m, including the main effects. If m is also graphical, $\mathcal{D}$ is the set of all non-empty complete subsets of the corresponding independence graph.

It follows from the theory of log-linear models (for example, see [22]) and from Lemma 3.1 that, for $E \notin \mathcal{D}$ or for $E \in \mathcal{D}$ but $i_E \notin \mathcal{I}_E^*$

$$\theta(i_E) = 0. \tag{6}$$

Therefore, for $i_E \in \mathcal{I}_E^*$, (3) becomes

$$\log p(i_E, i_{E^c}^*) = \log p(i(E)) = \theta_\emptyset + \sum_{F \subseteq E, F \in \mathcal{D}, i_F \in \mathcal{I}_F^*} \theta(i_F). \tag{7}$$

and after the change of variable $(n(i), i \in \mathcal{I}^*) \mapsto (n(i_E), E \in \mathcal{E}_\ominus)$, we obtain the following expression for the multinomial distribution associated with m.

Lemma 3.2. *The probability function of the multinomial distribution (1) corresponding to the model m can be represented as a natural exponential family with the marginal cell counts $y = (n(i_D), i_D \in \mathcal{I}_D^*, D \in \mathcal{D})$ as canonical statistics, and with density, with respect to the counting measure, proportional to*

$$f(y; \theta_\mathcal{D}, N) = \exp\left\{\sum_{D \in \mathcal{D}} \sum_{i_D \in \mathcal{I}_D^*} \theta(i_D) n(i_D) - N \log\Big(1 + \sum_{E \in \mathcal{E}_\ominus, i_E \in \mathcal{I}_E^*} \exp \sum_{F \subseteq_\mathcal{D} E} \theta(i_F)\Big)\right\}, \tag{8}$$

It is important to note that

$$\theta_\mathcal{D} = (\theta(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}_D^*), \tag{9}$$

is the canonical parameter and

$$p_\mathcal{D} = (p(i(D)), D \in \mathcal{D}, i_D \in \mathcal{I}_D^*), \tag{10}$$

is the cell probability parameter of this multinomial distribution. The remaining cell probabilities $p(i(E)), E \notin \mathcal{D}$ are not free and are a function of $p_\mathcal{D}$.

3.3. The Diaconis–Ylvisaker conjugate prior

The distribution of the marginal counts $Y = y = (n(i_D), i_D \in \mathcal{I}_D^*, D \in \mathcal{D})$ of a contingency table with cell counts $n(i), i \in \mathcal{I}$ as given in (8) is a natural exponential family. It follows immediately that the density of the conjugate prior for $\theta_{\mathcal{D}}$, with respect to the Lebesgue measure is

$$\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha) = I_{\mathcal{D}}(s,\alpha)^{-1} h(\theta_{\mathcal{D}}; s, \alpha), \tag{11}$$

where $I_{\mathcal{D}}(s,\alpha) = \int_{\mathbb{R}^{d_{\mathcal{D}}}} h(\theta_{\mathcal{D}}; s, \alpha) \mathrm{d}\theta_{\mathcal{D}}$ is the normalizing constant of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha)$, the dimension of the parameter space $d_{\mathcal{D}}$ is $\sum_{D\in\mathcal{D}} \prod_{\gamma\in D}(|I_\gamma| - 1)$ and

$$h(\theta_{\mathcal{D}}; s, \alpha) = \exp\left\{ \sum_{D\in\mathcal{D}} \sum_{i_D \in \mathcal{I}_D^*} \theta(i_D) s(i_D) - \alpha \log\Big(1 + \sum_{E\in\mathcal{E}_\ominus} \sum_{i_E\in\mathcal{I}_E^*} \exp \sum_{F\subseteq_{\mathcal{D}} E} \theta(i_F)\Big)\right\}. \tag{12}$$

The corresponding hyper-parameters are:

$$(s,\alpha) = (s(i_D), D \in \mathcal{D}, i_D \in \mathcal{I}_D^*, \alpha), \quad s \in \mathbb{R}^{d_{\mathcal{D}}},\ \alpha \in \mathbb{R}. \tag{13}$$

From Theorem 1 of [20] it follows that a necessary and sufficient condition for the distribution (11) to be proper (i.e., $I_{\mathcal{D}}(s,\alpha) < +\infty$) is that s represent the $\mathcal{D}$-marginal counts $s(i_D)$ of a contingency table (s) that has strictly positive real numbers $s(i), i \in \mathcal{I}$ as cell entries and that α is the grand total of (s), i.e. $\alpha = \sum_{i\in\mathcal{I}} s(i)$. Remark that $s(i)$ are not necessarily integers.

Given the prior $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha)$ and the multinomial likelihood expressed as a function of the marginal cell counts y as in (8), the corresponding posterior distribution of $\theta_{\mathcal{D}}$ is

$$\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s+y,\alpha+N) = I_{\mathcal{D}}(s+y,\alpha+N)^{-1} h(\theta_{\mathcal{D}}; s+y, \alpha+N).$$

Here $s+y = (s(i_D)+n(i_D), i_D \in \mathcal{I}_D^*, D \in \mathcal{D})$. We remark that $s(i_D)+n(i_D)$ represents the (i_D)-marginal count of the table $(s+n) = (s(i)+n(i), i \in \mathcal{I})$ obtained by augmenting the observed counts $n(i)$ with the prior cell entries $s(i)$. The grand total of this table is $\alpha + N$.

3.4. Finding the mode of the DY conjugate prior

The mode of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha)$ is given by

$$\widehat{\theta}_{\mathcal{D}} = \mathrm{argmax}_{\theta_{\mathcal{D}}\in\mathbb{R}^{d_{\mathcal{D}}}} h(\theta_{\mathcal{D}}; s, \alpha). \tag{14}$$

As shown in the proof of Lemma 3.2, we have that $h(\theta_{\mathcal{D}}; s, \alpha) = g((s), (p))$ where g is given by (1). Therefore (14) is equivalent to finding the maximum likelihood estimate of (p), the cell probabilities for the multinomial model m. Since all the cell entries in (s) are strictly positive, $g((s), (p))$ has a unique mode $(\widehat{p}) = (\widehat{p}(i), i \in \mathcal{I})$ that is identified using the iterative proportional fitting (IPF) algorithm – see [1,21]. We use (2) to obtain $\widehat{\theta}_{\mathcal{D}} = (\widehat{\theta}(i_D), i_D \in \mathcal{I}_D^*, D \in \mathcal{D})$ from $(\widehat{p})$.

The mode of the posterior distribution $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s+y,\alpha+N)$ can be computed in a similar manner. The posterior mode exists and is unique because $(s+n)$ has only strictly positive cell entries even if (n) has many counts of zero.

4. Computing marginal likelihoods

Let (n) be a contingency table and let (s,α) be hyper-parameters for the conjugate prior $\pi_{\mathcal{D}}(\theta|s,\alpha)$ associated with a hierarchical log-linear model m specified by the interactions $\mathcal{D}$. The marginal likelihood of m is the ratio of normalizing constants of the posterior and the prior for θ:

$$\Pr((n)|m) = I_{\mathcal{D}}(y+s, N+\alpha)/I_{\mathcal{D}}(s,\alpha).$$

Knowing how to efficiently evaluate the marginal likelihood of a model is key for the stochastic search methods discussed in this paper. We show how to calculate the normalizing constant $I_{\mathcal{D}}(s,\alpha)$ of the distribution $\pi_{\mathcal{D}}(\theta|s,\alpha)$ in (11) for hierarchical, graphical and decomposable log-linear models. The posterior normalizing constant $I_{\mathcal{D}}(y+s, N+\alpha)$ is computed in a similar manner.

4.1. Hierarchical log-linear models

In the most general case when m is a hierarchical log-linear model, we use the Laplace approximation [23] to estimate $I_{\mathcal{D}}(s,\alpha) = \int_{\mathbb{R}^{d_{\mathcal{D}}}} h_{s,\alpha}(\theta_{\mathcal{D}})\mathrm{d}\theta_{\mathcal{D}}$ where $h_{s,\alpha}(\theta_{\mathcal{D}}) = h(\theta_{\mathcal{D}}; s, \alpha)$. Let $\widehat{\theta}_{\mathcal{D}}$ be the mode of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s,\alpha)$ calculated using IPF as explained in Section 3.4. The Laplace approximation to $I_{\mathcal{D}}(s,\alpha)$ is

$$\begin{aligned}
\widehat{I_{\mathcal{D}}(s,\alpha)} &= \int_{\mathbb{R}^{d_{\mathcal{D}}}} \exp\left\{\log h_{s,\alpha}(\widehat{\theta}_{\mathcal{D}}) + \frac{1}{2}(\theta_{\mathcal{D}} - \widehat{\theta}_{\mathcal{D}})^t H_{s,\alpha}(\widehat{\theta}_{\mathcal{D}})(\theta_{\mathcal{D}} - \widehat{\theta}_{\mathcal{D}})\right\}\mathrm{d}\theta_{\mathcal{D}},\\
&\approx h_{s,\alpha}(\widehat{\theta}_{\mathcal{D}}) \int_{\mathbb{R}^{d_{\mathcal{D}}}} \exp\left\{\frac{1}{2}(\theta_{\mathcal{D}} - \widehat{\theta}_{\mathcal{D}})^t H_{s,\alpha}(\widehat{\theta}_{\mathcal{D}})(\theta_{\mathcal{D}} - \widehat{\theta}_{\mathcal{D}})\right\}\mathrm{d}\theta_{\mathcal{D}},\\
&\approx h_{s,\alpha}(\widehat{\theta}_{\mathcal{D}})(2\pi)^{\frac{d_{\mathcal{D}}}{2}} \det(H_{s,\alpha}(\widehat{\theta}_{\mathcal{D}}))^{-1/2},
\end{aligned}$$

where $(\theta_{\mathcal{D}} - \widehat{\theta}_{\mathcal{D}})$ is a $d^{\mathcal{D}}$-dimensional column vector and

$$H_{s,\alpha}(\widehat{\theta}_{\mathcal{D}}) = \frac{\mathrm{d}^2}{\mathrm{d}\theta_{\mathcal{D}}^2}\left\{\sum_{D\in\mathcal{D}}\sum_{i_D\in\mathcal{I}_D^*}\theta(i_D)s(i_D) - \alpha\log\Big(1 + \sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}_E^*}\exp\sum_{F\subseteq_{\mathcal{D}}E}\theta(i_F)\Big)\right\}\Bigg|_{\widehat{\theta}_{\mathcal{D}}}.$$

Let us compute the first derivative

$$\begin{aligned}
\frac{\mathrm{d}h_{s,\alpha}(\theta_{\mathcal{D}})}{\mathrm{d}\theta(i_D)} &= s(i_D) - \alpha\frac{\sum_{\substack{G\in\mathcal{E}_\ominus\\ G\supseteq D}}\sum_{\substack{j_G\in\mathcal{I}_G^*\\ (j_G)_D=i_D}}\exp\sum_{F\subseteq_{\mathcal{D}}E}\theta(i_F)}{1 + \sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}_E^*}\exp\sum_{F\subseteq_{\mathcal{D}}E}\theta(i_F)},\\
&= s(i_D) - \alpha\sum_{\substack{G\in\mathcal{E}_\ominus\\ G\supseteq D}}\sum_{\substack{j_G\in\mathcal{I}_G^*\\ (j_G)_D=i_D}}p(j_G).
\end{aligned}$$

Using the expression for $\frac{\mathrm{d}p(j(G))}{\mathrm{d}\theta(l_H)}$ derived in [15], we obtain

$$\begin{aligned}
\frac{\mathrm{d}^2h_{s,\alpha}(\theta_{\mathcal{D}})}{\mathrm{d}\theta(i_D)\mathrm{d}\theta(l_H)} &= -\alpha\sum_{\substack{G\in\mathcal{E}_\ominus\\ G\supseteq D}}\sum_{\substack{j_G\in\mathcal{I}_G^*\\ (j_G)_D=i_D}}\frac{\mathrm{d}p(j(G))}{\mathrm{d}\theta(l_H)},\\
&= -\alpha\sum_{\substack{G\in\mathcal{E}_\ominus\\ G\supseteq D}}\sum_{\substack{j_G\in\mathcal{I}_G^*\\ (j_G)_D=i_D}}p(j(G))\left[\delta_{(j_G)_H}(l_H) - \sum_{\substack{(j_C)_H=l_H\\ C\in\mathcal{E}_\ominus, j_C\in\mathcal{I}_C^*}}p(j(C))\right]
\end{aligned}$$

where

$$\delta_{(j_G)_H}(l_H) = \begin{cases}1, & \text{if } (j_G)_H = l_H,\\ 0, & \text{otherwise.}\end{cases}$$

For binary data, this yields

$$\frac{\mathrm{d}^2h_{s,\alpha}(\theta_{\mathcal{D}})}{\mathrm{d}\theta(D)\mathrm{d}\theta(H)} = -\alpha\sum_{G\supseteq D}p(G)\left[\delta_{\supseteq H}(G) - \sum_{C\supseteq H}p(C)\right],$$

where

$$\delta_{\supseteq H}(G) = \begin{cases}1, & \text{if } G\supseteq H,\\ 0, & \text{otherwise.}\end{cases}$$

The Hessian matrix $H_{s,\alpha}(\widehat{\theta}_{\mathcal{D}})$ is therefore the $d_{\mathcal{D}} \times d_{\mathcal{D}}$ matrix with (i_D, l_H) entries, $D \in \mathcal{D}$, $i_D \in \mathcal{I}_D^*$, $H \in \mathcal{D}$, $l_H \in \mathcal{I}_H^*$ given by

$$-\alpha \sum_{\substack{G \in \mathcal{E}_{\ominus} \\ G \supseteq D}} \sum_{\substack{j_G \in \mathcal{I}_G^* \\ (j_G)_D = i_D}} p(j(G)) \left[\delta_{(j_G)_H}(l_H) - \sum_{\substack{(j_C)_H = l_H \\ C \in \mathcal{E}_{\ominus}, j_C \in \mathcal{I}_C^*}} p(j(C)) \right].$$

4.2. Graphical log-linear models

Let us assume that the log-linear model m is Markov with respect to an arbitrary undirected graph G. We develop a more efficient way of approximating $I_{\mathcal{D}}(s, \alpha)$ based on the strong hyper-Markov property [17] of the generalized hyper-Dirichlet $\pi_{\mathcal{D}}(\theta|s, \alpha)$.

Let $P_1, \ldots, P_k$ be a perfect sequence of the prime components of G and let $S_2, \ldots, S_k$ be the corresponding separators, where $S_l = \left(\cup_{j=1}^{l-1} P_j\right) \cap P_l$, $l = 2, \ldots, k$. We use the notation $\mathcal{D}^{P_l}$ ($l = 1, \ldots, k$) and $\mathcal{D}^{S_l}$ ($l = 2, \ldots, k$) for the collection of complete subsets of the induced sub-graphs G_{P_l} and G_{S_l}, respectively. More precisely, $\mathcal{D}^A$ for some $A \subset V$ defines the graphical log-linear model for the A-marginal of (n) with independence graph G_A, the subgraph of G induced by A. The parameters of the P_l-marginal and the S_l-marginal multinomials are $\theta(\mathcal{D}^{P_l})$ and $\theta(\mathcal{D}^{S_l})$, respectively. Massam et al. [15] prove that $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$ is strong hyper-Markov with respect to G and can be written as a hyper-Markov combination of the marginal distribution of $\theta(\mathcal{D}^{P_l})$ and $\theta(\mathcal{D}^{S_l})$. This implies that the normalizing constant $I_{\mathcal{D}}(s, \alpha)$ of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$ is equal to

$$I_{\mathcal{D}}(s, \alpha) = \frac{\prod_{l=1}^{k} I_{\mathcal{D}^{P_l}}(s^{P_l}, \alpha)}{\prod_{l=2}^{k} I_{\mathcal{D}^{S_l}}(s^{S_l}, \alpha)}, \tag{15}$$

i.e., it is the Markov ratio of normalizing constants for the lower-dimensional models $\mathcal{D}^{P_l}, l = 1, \ldots, k$ and $\mathcal{D}^{S_l}$, $l = 2, \ldots, k$ Markov with respect to the prime components and the separators of the graph G.

If A is a prime component of G, G_A is might not be complete and we need to use the Laplace approximation from Section 4.1 to calculate the normalizing constant $I_{\mathcal{D}^A}(s^A, \alpha)$. On the other hand, if G_A is complete, no approximation is needed because the normalizing constant is that of a Dirichlet. More precisely, we have [15]:

$$I_{\mathcal{D}^A}(s^A, \alpha) = \frac{\Gamma(\alpha^A_{\emptyset})}{\Gamma(\alpha)} \prod_{D \in \mathcal{D}^A, i_D \in \mathcal{I}_D^*} \Gamma(\alpha^A(i_D, i^*_{A \setminus D})), \tag{16}$$

where

$$\alpha^A(i_D, i^*_{A \setminus D}) = \sum_{A \supseteq F \supseteq D} \sum_{\substack{j_F \in \mathcal{I}_F^* \\ (j_F)_D = i_D}} (-1)^{|F \setminus D|} s(j_F),$$

$$\alpha^A_{\emptyset} = \alpha + \sum_{D \subseteq A} (-1)^{|D|} \sum_{i \in \mathcal{I}_D^*} s(i_D).$$

If A is a separator of G the subgraph G_A is always complete and we can use (16).

Although the IPF algorithm can efficiently determine the mode of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$, it can still be slow for large, sparse contingency tables since it has to take into consideration every single cell. The divide-and-conquer method for estimating $I_{\mathcal{D}}(s, \alpha)$ based on the sequence of prime components and separators of the independence graph is likely to be faster than the Laplace approximation from Section 4.1 since it breaks the original table into smaller-dimensional marginals whose corresponding normalizing constants can be calculated in parallel.

4.3. Decomposable log-linear models

We further assume that the log-linear model m is Markov with respect to a decomposable undirected graph G. A graph is decomposable if and only if each of its prime components is complete [24]. Assume that G is decomposed into the complete prime components $P_1, \ldots, P_k$ and the sequence of separators $S_2, \ldots, S_k$. Then $I_{\mathcal{D}}(s, \alpha)$ is calculated using formula (15) with each $I_{\mathcal{D}^A}(s^A, \alpha)$ for $A \in \{P_1, \ldots, P_k, S_2, \ldots, S_k\}$ given by (16). Therefore the normalizing constant for a decomposable log-linear model can be calculated exactly since $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$ is hyper-Dirichlet [15].

5. The mode oriented stochastic search (MOSS) algorithm

The Bayesian paradigm for model determination involves choosing models with high posterior probability selected from a set $\mathcal{M}$ of competing models. Godsill [25] provides an excellent review of MCMC methods for exploring $\mathcal{M}$ such as the reversible jump sampler of Green [26] or the product space scheme of Carlin and Chib [27]. The number of iterations required to achieve convergence can increase rapidly if the Markov chain is run over the product space of $\mathcal{M}$ and the corresponding model parameters, due to the high dimensionality of the state space. For this reason there has been a recent development of stochastic search methods in which the model parameters are integrated out. Examples of such methods are the Markov chain Monte Carlo model composition (MC^3) algorithm of Madigan and York [9] and the shotgun stochastic search (SSS) algorithm of Jones et al. [13] and Hans et al. [14]. If the posterior probability of a model is readily available from its marginal likelihood, up to the normalizing constant

$$\left[\sum_{m\in\mathcal{M}} \Pr(m|(n))\right]^{-1}, \tag{17}$$

there is no substantive need to sample from the whole posterior distribution $\{\Pr(m|(n)) : m \in \mathcal{M}\}$. A stochastic search method is designed to visit regions of high posterior probability and is not constrained to be a Markov chain on $\mathcal{M}$. Jones et al. [13] and Hans et al. [14] showed that SSS consistently finds better models faster than MC^3 for linear regression and Gaussian graphical models.

In this section we further exploit the principles behind SSS and propose a novel stochastic search method which we call the mode oriented stochastic search (MOSS, henceforth). MOSS focuses on determining the set of models

$$\mathcal{M}(c) = \left\{m \in \mathcal{M} : \Pr(m|(n)) \geq c \cdot \max_{m'\in\mathcal{M}} \Pr(m'|(n))\right\}, \tag{18}$$

where $c \in (0, 1)$ and (n) is the data. We follow Occam's window idea of Madigan and Raftery [8] and discard models with a low posterior probability compared to the highest probability model. Raftery et al. [28] described an MCMC approach to identify models in $\mathcal{M}(c)$ for linear regression.

In order to implement MOSS we need to compute the posterior probability $\Pr(m|(n)) \propto \Pr((n)|m)\Pr(m)$ of any given model $m \in \mathcal{M}$. In Section 4 we showed how to evaluate the marginal likelihood $\Pr((n)|m)$ for decomposable, graphical and arbitrary log-linear models. Throughout this paper we assume that the models in $\mathcal{M}$ are equally likely a priori, so that $\Pr(m|(n)) \propto \Pr((n)|m)$. The determination of the normalizing constant (17) is not required in our framework.

We also need a way to traverse the space $\mathcal{M}$. To this end, we associate with each candidate model $m \in \mathcal{M}$ a neighborhood $\mathrm{nbd}(m) \subset \mathcal{M}$. Any two models in $m, m' \in \mathcal{M}$ are connected through at least a path $m = m_1, m_2, \ldots, m_k = m'$ such that $m_j \in \mathrm{nbd}(m_{j-1})$ for $j = 2, \ldots, k$. The neighborhoods are defined with respect to the class of models considered:

(a) *Hierarchical log-linear models.* The neighborhood of a hierarchical model m consists of those hierarchical models obtained from m by adding one of its dual generators (i.e., minimal interaction terms not present in the model) or deleting one of its generators (i.e., maximal interaction terms present in the model). For details see [2] and [7].

(b) *Graphical log-linear models.* The neighborhood of a graphical model m with independence graph G is defined by the graphs obtained by adding or removing one edge from G. The size of the neighborhoods is therefore constant.

(c) *Decomposable log-linear models*. Here the neighborhood of a model is obtained by adding or deleting edges such that the resulting graph is still decomposable — see [17] or [11] for details. The sizes of the neighborhoods of two decomposable graphs are not necessarily the same even if they differ by exactly one edge.

To implement MOSS, we need a current list $\mathcal{S}$ of models that is updated during the search. We define the subset $\mathcal{S}(c)$ of $\mathcal{S}$ in the same way we defined $\mathcal{M}(c)$ based on $\mathcal{M}$. In order to allow our search to escape local optima by occasionally moving to models with lower posterior probability and exploring their neighborhoods, we define $\mathcal{S}(c')$ with $0 < c' \leq c$ so that $\mathcal{S}(c) \subseteq \mathcal{S}(c')$. We also need to choose the probability q of pruning the models in $\mathcal{S} \setminus \mathcal{S}(c)$. A model m is called *explored* if all its neighbors $m' \in \text{nbd}(m)$ have been visited. A model in $\mathcal{S}$ can be explored or unexplored. MOSS proceeds as follows:

PROCEDURE MOSS(c, c', q)

(a) Initialize the starting list of models $\mathcal{S}$. For each model $m \in \mathcal{S}$, calculate its posterior probability $\Pr(m|(n))$ up to the normalizing constant (17) and record it. Mark m as unexplored.
(b) Let $\mathcal{L}$ be the set of unexplored models in $\mathcal{S}$. Sample a model $m \in \mathcal{L}$ according to probabilities proportional to $\Pr(m|(n))$ normalized within $\mathcal{L}$. Mark m as explored.
(c) For each $m' \in \text{nbd}(m)$, check if m' is currently in $\mathcal{S}$. If it is not, calculate its posterior probability $\Pr(m'|(n))$ up to the normalizing constant (17) and record it. If $m' \in \mathcal{S}(c')$, include m' in $\mathcal{S}$ and mark m' as unexplored. If m' is the model with the highest posterior probability in $\mathcal{S}$, eliminate from $\mathcal{S}$ the models in $\mathcal{S} \setminus \mathcal{S}(c')$.
(d) With probability q, eliminate from $\mathcal{S}$ the models in $\mathcal{S} \setminus \mathcal{S}(c)$.
(e) If all the models in $\mathcal{S}$ are explored, eliminate from $\mathcal{S}$ the models in $\mathcal{S} \setminus \mathcal{S}(c)$ and STOP. Otherwise go back to step (b).

END.

We output $\mathcal{S} = \mathcal{S}(c)$ and further use it to quantify the uncertainty related to our model choice. Kass and Raftery [29] suggest that choosing c in one of the intervals $(0, 0.01]$, $(0.01, 0.1]$, $(0.1, 1/3.2]$, $(1/3.2, 1]$ is equivalent to, respectively, discarding models with decisive, strong, substantial or "not worth more than a bare mention" evidence against them with respect to m_h. The number of models in $\mathcal{M}(c)$ increases as c decreases, so $\mathcal{M}(c)$ can be exhaustively enumerated for higher values of c. We note that producing the entire set $\mathcal{M}$ is not practically possible for the example we analyze in this paper.

The choice of the other two parameters of the MOSS algorithm is merely a way to balance the computing time required by the procedure and the computing resources available with its ultimate successful identification of $\mathcal{M}(c)$. If c' is set to be too close to c, MOSS might end before reaching m_h due to its inability to escape local modes. On the other hand, setting c' to an extremely low value could mean that MOSS might take a long time to end since the neighborhoods of too many models would have to be explored. In addition, managing the list $\mathcal{L}$ might become cumbersome due to its size. Larger values of q decrease the number of iterations until MOSS ends since models with lower posterior probability are more often discarded from $\mathcal{S}$. However, these models might be on paths between $\mathcal{S}$ and m_h, hence MOSS could end before identifying m_h if these paths are broken.

In our experience finding suitable values for the parameters c' and q has been far less burdensome than calibrating the number of iterations needed by a Markov chain to find the best models in $\mathcal{M}$. We remark that there is a rich literature dedicated to assessing the convergence of MCMC algorithms to their stationary distributions — see, for example, [30]. To the best of our knowledge, there is no rigorous approach for establishing whether an MCMC algorithm has actually found models in $\mathcal{M}(c)$. We suggest running MOSS several times to make sure that the same final set of models has been reached. We also recommend using values of c' and q as small as possible in order to visit as many models as possible. In fact, we view any MOSS(c, c', q) procedure with $c' > 0$ and $q > 0$ as an approximation to the MOSS($c, 0, 0$) procedure. In the limiting case when $c' = q = 0$, MOSS always outputs $\mathcal{M}(c)$ as we prove below.

Proposition 5.1. *MOSS(c, 0, 0) visits the entire set of candidate models $\mathcal{M}$.*

Proof. Let m_0 be a model included in the starting list of models from step (a) of the algorithm. Let $m_1, \ldots, m_k, m_{k+1} = m$ be a path that connects m_0 with an arbitrary model $m \in \mathcal{M}$, i.e. $m_j \in \text{nbd}(m_{j-1})$ for $j = 1, \ldots, k+1$. For l equal to, successively, 0, 1, 2, $\ldots$, k, let us assume that at the current iteration $m_l \in \mathcal{S}$, and $m_{l+1} \notin \mathcal{S}$. We want to show that MOSS must include m_{l+1} in $\mathcal{S}$ before it ends. Since $m_{l+1} \in \text{nbd}(m_l)$, m_l is still unexplored, i.e. $m_l \in \mathcal{L}$. The probability that m_l is selected at step (b) of the procedure is therefore:

$$\Pr(m_l|(n))/\left[\sum_{m' \in \mathcal{L}} \Pr(m'|(n))\right]. \tag{19}$$

In the worst possible case, MOSS explores all the other models in $\mathcal{L}$ before m_l but because c' and q are both equal to 0, m_l remains in $\mathcal{L}$ and MOSS cannot end before $\mathcal{L}$ becomes empty — see step (e) of MOSS. Since, in this worst possible case, m_l is then the only model in $\mathcal{L}$, the probability (19) is equal to 1. Hence MOSS selects m_l and visits all its neighbors. This implies that m_{l+1} is included in $\mathcal{S}$, which in turn implies that MOSS reaches m starting from m_0.

This result shows that MOSS$(c, 0, 0)$ ends only when the entire $\mathcal{M}$ has been explored, hence $\mathcal{S} = \mathcal{M}$ at step (e) of the last iteration of the procedure. MOSS$(c, 0, 0)$ includes in $\mathcal{S}$ every model it visits and never discards any of these models. This implies that MOSS$(c, 0, 0)$ explores every model in $\mathcal{M}$ exactly once. By comparison, the procedure MOSS(c, c, q), $q \in [0, 1]$, discards every model in $\mathcal{S} \setminus \mathcal{S}(c)$ so that $\mathcal{S} = \mathcal{S}(c)$ at all times. Therefore MOSS(c, c, q) might not identify a model $m \in \mathcal{M}(c)$ if lower posterior probability models in $\mathcal{S} \setminus \mathcal{S}(c)$ are needed to connect the starting set of models from step (a) of MOSS to m.

MOSS never discards a model in $\mathcal{M}(c)$ from the current set of models $\mathcal{S}$ for any choices of c' and q. This means that the models in $\mathcal{M}(c)$ are never explored twice during a run of the procedure. On the other hand, MOSS might explore models in $\mathcal{M} \setminus \mathcal{M}(c)$ more than once if $c' \in (0, c)$ and $q > 0$. MCMC algorithms can revisit all the models in $\mathcal{M}$ indefinitely. In an MCMC search, the next model to be explored is selected only from the neighbors of the model evaluated in the previous iteration. In a MOSS search, this model is selected from the most promising models identified so far. Models with higher posterior probability are more likely to be selected for exploration than models with lower posterior probability. □

5.1. Household study in Rochdale: Revisited

We use MOSS to analyze the Rochdale data. Whittaker [4] pointed out that the severe imbalance in the cell counts of this sparse eight-way table is often found in social survey analysis. Whittaker's analysis was based on the assumption that models with higher-order interactions cannot be fit to this data due to the zero counts in the marginals that in turn translate into the non-existence of MLEs and into difficulties in correctly calculating the number of degrees of freedom. Whittaker starts with all two-way interaction models and sequentially eliminates edges based on their deviances. All the higher-order interactions were discarded up front. Whittaker chooses the model

$$fg|ef|dh|dg|cg|cf|ce|bh|be|bd|ag|ae|ad|ac. \tag{20}$$

To the best of the authors' knowledge, there has been no other published analysis of this dataset following Whittaker's work. We present a new analysis of this data that confirms Whittaker's intuition but also reveals that there actually exists a three-way interaction *bdh* that is supported by the data. This interaction indicates a strong connection between wife's age, her child's age and the presence of another working member in the family.

We penalize for model complexity by choosing $\alpha = 1$ in the specification of the conjugate prior. This means that we augment the actual data with small fictive counts of 2^{-8}. We run five replicates of MOSS within the space of decomposable, graphical and hierarchical log-linear models. The search over decomposable models was done with $c = 0.1$, $c' = 10^{-5}$ and $q = 0.001$. We increased the pruning probability to 0.1 for the graphical and hierarchical searches due to the larger number of models that had to be kept in the list $\mathcal{S}$. The search over decomposable models was started from random starting models. The graphical model search was started from the top decomposable models

Table 1
The models with the highest posterior probabilities identified by MOSS for the Rochdale data. We report the models whose normalized posterior probabilities are greater than 0.05. We also give the minimum, median and maximum number of models visited by MOSS before completion across the five search replicates.

Search	Top models		Models evaluated
Dec.	*efg|beg|bdh|bdg|adg|acg*	0.436	1123|5608|6240
	efg|ceg|bdh|adg|acg	0.369	
	efg|ceg|beg|bdh|bdg|acg	0.069	
	efg|bh|beg|bdg|adg|acg	0.068	
	efg|ceg|bh|bd|adg|acg	0.058	
	efg|beg|bdh|bdg|adg|acg	med.	
Graph./PM	*fg|ef|be|bdh|bdg|adg|acg|ace*	0.462	240|369|608
	fg|ef|bh|be|bd|adg|acg|ace	0.337	
	fg|ef|bh|be|bdg|adg|acg|ace	0.072	
	fg|ef|ce|be|bdh|bdg|adg|acg	0.067	
	fg|ef|ce|bh|be|bd|adg|acg	0.061	
	fg|ef|be|bdh|bdg|adg|acg|ace	med.	
Graph./Lapl	*fg|ef|be|bdh|adg|acg|ace*	0.507	290|515|926
	fg|ef|ce|be|bdh|adg|acg	0.184	
	efg|ceg|be|bdh|adg|acg	0.112	
	fh|fg|ef|be|bdh|adg|acg|ace	0.087	
	fg|ef|bg|be|bdh|ad|acg|ace	0.056	
	fg|ef|be|bdh|bdg|adg|acg|ace	0.055	
	fg|ef|be|bdh|adg|acg|ace	med.	
Hierar.	*fg|ef|dg|cg|cf|ce|be|bdh|ag|ae|ad|ac*	0.076	1391|1417|1617
	fg|ef|dg|cg|ce|be|bdh|ag|ae|ad|ac	0.069	
	fg|ef|dg|cf|ce|be|bdh|ae|ad|acg	0.057	
	fg|ef|dg|ce|be|bdh|ae|ad|acg	0.052	
	fg|ef|dg|cg|cf|ce|be|bdh|ag|ae|ad|ac	med.	

identified by MOSS, while the hierarchical model search was started from the top graphical models identified. Replacing the random starting models with a set of models that are known to give a fairly good representation of the data leads to a more efficient stochastic search that visits a smaller number of models.

Table 1 shows the top decomposable, graphical and hierarchical log-linear models identified by MOSS. Remark the similarity of the models obtained by estimating the marginal likelihoods of graphical models by a single Laplace approximation or by decomposing the independence graph in its prime components. The hierarchical log-linear model with the highest posterior probability differs by only one interaction term *bdh* from the model proposed by Whittaker.

Table 1 also gives the number of models evaluated by MOSS before its completion. About 5600 models had to be examined in the decomposable case. Evaluating the marginal likelihood of a decomposable model is efficient since explicit formulas exist. Since numerical approximations to marginal likelihoods have to be used in the graphical and hierarchical cases, the number of models visited should be as small as possible because of the increased computing time needed to evaluate each model. Fewer graphical and hierarchical models were evaluated by MOSS because the search was started from models that were not far from the highest probable models in each class. MOSS determined the top graphical models out of 2^{28} possible graphs by visiting less than one thousand models. MOSS seems to work very well for hierarchical log-linear models by identifying the top models out of 5.6×10^{22} possible hierarchical log-linear models [7] by visiting less than 2000 models.

6. The Bayesian iterative proportional fitting algorithm

Consider a hierarchical log-linear model with an irreducible generating class $\mathcal{A} = \{A_i, i = 1, \ldots, k\}$ and with constraints $\mathcal{D}$ defined as the set of subsets of A_i, $i = 1, \ldots, k$. Finding the mode of the posterior distribution $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s + y, \alpha + N)$ or the mode of the prior distribution $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$ can be done in a computationally efficient manner using the IPF algorithm — see Section 3.4. Although

this solves the problem of fitting log-linear models, it is important to know how to sample from these constrained distributions in order to quantify estimation uncertainty and to produce Bayesian estimates of other quantities of interest that are nonlinear transformations of $\theta_{\mathcal{D}}$.

To this end, Gelman et al. [31] and Schafer [32] proposed the Bayesian iterative proportional fitting algorithm for simulating random draws from the constrained Dirichlet posterior for a given log-linear model. The Bayesian IPF is similar to the classical IPF algorithm, except that sequentially updating the parameters θ based on each fixed marginal is replaced with an adjustment based on a marginal table with the same structure whose entries have been drawn from Gamma distributions. Piccioni [16] exploits the theory of regular exponential families with cuts to formally construct a Gibbs sampler algorithm for sampling from their natural conjugate densities. Asci and Piccioni [33] give an extension to improper target distributions.

In this section we generalize to arbitrary contingency tables the version of Bayesian IPF for binary data described in [33]. The algorithm starts with a random set of $\theta^{(0)}_{\mathcal{D}} = (\theta^{(0)}(i_D), i_D \in \mathcal{I}^*_D, D \in \mathcal{D})$ that can be generated, for example, from independent standard normal distributions. The remaining elements of $\theta^{(0)} \in \mathbb{R}^{|\mathcal{E}|} = \mathbb{R}^{2^{|V|}}$ are set to zero, i.e. $\theta^{(0)}(i_E) = 0$ for $E \notin \mathcal{D}$ or $E \in \mathcal{D}$, $i_E \notin \mathcal{I}^*_E$. A cycle of Bayesian IPF sequentially goes through each sufficient configuration A_l, $l = 1, \ldots, k$ and updates the current sampled values $\theta^{(\text{old})}$ to a new set of sampled values $\theta^{(\text{new})}$ in the following way:

(a) Generate independent gamma variables for the marginal expected cell counts $\tau^{A_l}(i_D, i^*_{A_l\setminus D})$, $D \subseteq A_l$, $i_D \in \mathcal{I}^*_D \cup (i^*)_{A_l}$ according to the law

$$g_{A_l}(\tau^{A_l}(i_D, i^*_{A_l\setminus D})) \propto \tau^{A_l}(i_D, i^*_{A_l\setminus D})^{\alpha^{A_l}(i_D, i^*_{A_l\setminus D})-1} \exp(-\alpha\tau^{A_l}(i_D, i^*_{A_l\setminus D})),$$

where for $D \neq \emptyset$,

$$\alpha^{A_l}(i_D, i^*_{A_l\setminus D}) = \sum_{A_l \supseteq F \supseteq D} \sum_{\substack{j_F \in \mathcal{I}^*_F \\ (j_F)_D = i_D}} (-1)^{|F\setminus D|} s(j_F),$$

and

$$\alpha^{A_l}(i^*_{A_l}) = \alpha^{A_l}_{\emptyset} = \alpha + \sum_{D \subseteq A_l} (-1)^{|D|} \sum_{i_D \in \mathcal{I}^*_D} s(i_D).$$

In other words, generate independent gamma variables with shape parameter $\alpha^{A_l}(i_D, i^*_{A_l\setminus D})$ and scale parameter $1/\alpha$.

(b) Normalize the table obtained in (a) to obtain the table of A_l-*marginal* probabilities with entries

$$p^{A_l}(i_D, i^*_{A_l\setminus D}) = \frac{\tau^{A_l}(i_D, i^*_{A_l\setminus D})}{\sum_{F \subseteq A_l, i_F \in \mathcal{I}^*_F \cup (i^*)_{A_l}} \tau^{A_l}(i_F, i^*_{A_l\setminus F})}, \quad D \subseteq A_l,\ i_D \in \mathcal{I}^*_D \cup (i^*)_{A_l}.$$

(c) Compute the "marginal" $\theta^{A_l}(i_E)$, $E \subseteq A_l$, $i_E \in \mathcal{I}^*_E$ using the formula

$$\theta^{A_l}(i_E) = \log \prod_{F \subseteq E} p^{A_l}(i_F, i^*_{A_l\setminus F})^{(-1)^{|E\setminus F|}}. \tag{21}$$

(d) • For $E \in \mathcal{D}$, $E \subseteq A_l$, $i_E \in \mathcal{I}^*_E$, set $\theta^{(\text{new})}(i_E)$ to be equal to

$$\theta^{A_l}(i_E) + \sum_{F \subseteq E} (-1)^{|E\setminus F|-1} \log\Big(1 + \sum_{L \subseteq_\ominus A^c_l} \exp \sum_{\substack{H \subseteq F, G \subseteq_\ominus L \\ j_G \in \mathcal{I}^*_G}} \theta^{(\text{old})}(i_H, j_G)\Big).$$

• For $E \in \mathcal{D}$, $E \not\subseteq A_l$, set $\theta^{(\text{new})}(i_E) = \theta^{(\text{old})}(i_E)$.
• For $E \notin \mathcal{D}$ or $E \in \mathcal{D}$, $i_E \notin \mathcal{I}^*_E$, set $\theta^{(\text{new})}(i_E) = 0$.

Example. We consider the problem of predicting wife's economic activity a in the Rochdale data. Whittaker [4] page 285 considers the log-linear model $ac|ad|ae|ag$ induced by the generators of (20)

that involve a. Using maximum likelihood estimation of log-linear parameters in this model, he obtains the following estimates of the logistic regression of a on c, d, e and g:

$$\log \frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} = \text{const.} - 1.33c - 1.32d + 0.69e - 2.17g. \tag{22}$$

The corresponding standard errors of the regression coefficients are 0.3, 0.21, 0.2, 0.47. The generators involving a in the top hierarchical model identified by MOSS (see Table 1)

$$fg|ef|dg|cg|cf|ce|be|bdh|ag|ae|ad|ac \tag{23}$$

are again ac, ad, ae and ag which yield the regression equation

$$\log \frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} = \theta(a) + \theta(ac) + \theta(ad) + \theta(ae) + \theta(ag). \tag{24}$$

Using Bayesian IPF to produce 10,000 draws from the posterior probability associated with the log-linear model (23), we estimate the regression equation (24) to be:

$$\log \frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} = \text{const.} - 1.30c - 1.26d + 0.70e - 2.31g,$$

with standard errors 0.29, 0.2, 0.19 and 0.47, respectively. While these coefficient estimates are very close to Whittaker's estimates in (22), there is a major difference between how these estimates were obtained. We used the information in the full eight-way table to fit the log-linear model (23), while Whittaker used the five-way marginal associated with a, c, d, e and g to fit the log-linear model $ac|ad|ae|ag$.

7. Conclusions

In this paper we have developed MOSS which combines a new stochastic search algorithm for log-linear models with the conjugate prior for log-linear parameters of Massam et al. [15]. We showed that MOSS is a powerful technique to analyze multi-way contingency tables. Since we are able to integrate out the model parameters and compute marginal likelihoods, we avoid using MCMC techniques. MOSS is able to reach relevant log-linear models fast by evaluating a reduced set of models. Since models in each neighborhood can be evaluated in parallel, MOSS can be made considerably faster in a parallel implementation that takes advantage of cluster computing.

Penalizing for increased model complexity is immediate in this framework and is key in the analysis of sparse categorical data. The Bayesian IPF plays a crucial role in fitting log-linear models as well as the corresponding regressions based on these priors.

C++ code implementing various versions of MOSS for discrete data has been developed by the authors and can be downloaded from http://www.stat.washington.edu/adobra/software/mosstables/ This software is written only for dichotomous contingency tables. The methods presented in this paper hold for arbitrary multi-way cross-classifications and our code could therefore be extended to polychotomous data in a straightforward manner.

References

[1] Y.M.M. Bishop, S.E. Fienberg, P.W. Holland, Discrete Multivariate Analysis: Theory and Practice, M.I.T. Press, Cambridge, MA, 1975.
[2] D.E. Edwards, T. Havranek, A fast procedure for model search in multidimensional contingency tables, Biometrika 72 (1985) 339–351.
[3] A. Agresti, Categorical Data Analysis, Wiley, 1990.
[4] J. Whittaker, Graphical Models in Applied Multivariate Statistics, John Wiley & Sons, 1990.
[5] S.E. Fienberg, A. Rinaldo, Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation, J. Statist. Plann. Inference 137 (2007) 3430–3445.
[6] M. Clyde, E.I. George, Model uncertainty. Statist. Sci. 19 (2004) 81–94.
[7] P. Dellaportas, J.J. Forster, Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models, Biometrika 86 (1999) 615–633.
[8] D. Madigan, A. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window, J. Amer. Statist. Assoc. 89 (1994) 1535–1546.

[9] D. Madigan, J. York, Bayesian graphical models for discrete data, International Statistical Review 63 (1995) 215–232.
[10] D. Madigan, J. York, Bayesian methods for estimation of the size of a closed population, Biometrika 84 (1997) 19–31.
[11] C. Tarantola, MCMC model determination for discrete graphical models, Stat. Modelling 4 (2004) 39–61.
[12] P. Dellaportas, C. Tarantola, Model determination for categorical data with factor level merging, J. R. Stat. Soc. Ser. B 67 (2005) 269–283.
[13] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, M. West, Experiments in stochastic computation for high-dimensional graphical models, Statist. Sci. 20 (2005) 388–400.
[14] C. Hans, A. Dobra, M. West, Shotgun stochastic search for "large p" regression, J. Amer. Statist. Assoc. 102 (2007) 507–516.
[15] H. Massam, J. Liu, A. Dobra, A conjugate prior for discrete hierarchical log-linear models, Ann. Statist. (2009).
[16] M. Piccioni, Independence structure of natural conjugate densities to exponential families and the Gibbs sampler, Scand. J. Statist. 27 (2000) 111–127.
[17] A.P. Dawid, S.L. Lauritzen, Hyper Markov laws in the statistical analysis of decomposable graphical models, Ann. Statist. 21 (1993) 1272–1317.
[18] M. Knuiman, T. Speed, Incorporating prior information into the analysis of contingency tables, Biometrics 44 (1988) 1061–1071.
[19] R. King, S.P. Brooks, Prior induction for log-linear models for general contingency table analysis, Ann. Statist. 29 (2001) 715–747.
[20] P. Diaconis, D. Ylvisaker, Conjugate priors for exponential families, Ann. Statist. 7 (1979) 269–281.
[21] S.L. Lauritzen, Graphical Models, Clarendon Press, Oxford, 1996.
[22] J. Darroch, T. Speed, Additive and multiplicative models and interaction, Ann. Statist. 11 (1983) 724–738.
[23] L. Tierney, J. Kadane, Accurate approximations for posterior moments and marginal densities, J. Amer. Statist. Assoc. 81 (1986) 82–86.
[24] A. Dobra, S.E. Fienberg, Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, Proc. Natl. Acad. Sci. 97 (2000) 11185–11192.
[25] S.J. Godsill, On the relationship between Markov chain Monte Carlo methods for model uncertainty, J. Comput. Graph. Statist. 10 (2001) 1–19.
[26] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (1995) 711–732.
[27] B.P. Carlin, S. Chib, Bayesian model choice via Markov chain Monte Carlo, J. R. Stat. Soc. Ser. B 57 (1995) 473–484.
[28] A.E. Raftery, D. Madigan, J.A. Hoeting, Bayesian model averaging for linear regression models, J. Amer. Statist. Assoc. 92 (1997) 179–191.
[29] R.E. Kass, A.E. Raftery, Bayes factors, J. Amer. Statist. Assoc. 90 (1995) 773–795.
[30] C.P. Robert, Discretization and MCMC Convergence Assessment, in: Lecture Notes in Statistics, vol. 135, Springer-Verlag, 1998.
[31] A. Gelman, J.B. Carlin, H.S. Stern, D. Rubin, Bayesian Data Analysis, second ed., in: Texts in Statistical Science Series, Chapman & Hall, 2004.
[32] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.
[33] C. Asci, M. Piccioni, Functionally compatible local characteristics for the local specification of priors in graphical models, Scand. J. Statist. 34 (2007) 829–840.