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

Adrian Dobra

University of Washington, Seattle, USA

and Hélène Massam

York University, Toronto, Canada.

Summary. 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 also present a clustering algorithm for discrete data and develop regressions derived from log-linear models. We compare our model determination approach with similar results based on multivariate normal priors for log-linear parameters and Markov chain Monte Carlo stochastic search algorithms. The examples concern six-way, eight-way and sixteen-way contingency tables.

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

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 (Bishop et al., 1975) are the most suitable 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 (Edwards and Havranek, 1985; Agresti, 1990; Whittaker, 1990). 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 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 Fienberg and Rinaldo (2007) for an excellent discussion.

The Bayesian paradigm to model selection avoids these issues through the specification of prior distributions for model parameters (Clyde and George, 2004). Markov chain Monte

Address for correspondence: Adrian Dobra, Department of Statistics, University of Washington, Seattle, WA 98155-4322, USA.
E-mail: adobra@u.washington.edu

Carlo (MCMC) algorithms have been traditionally used to identify models with high posterior probability. Dellaportas and Forster (1999) 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 are, for example, Madigan and Raftery (1994); Madigan and York (1995, 1997); Tarantola (2004); Dellaportas and Tarantola (2005).

While MCMC methods seem to work well for problems involving a relatively small number of candidates models, they tend to be less efficient as the dimensionality of the model space grows exponentially. Jones et al. (2005) and Hans et al. (2007) 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 aggresively moving towards regions of high posterior probability in the models 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. 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 Massam et al. (2008). In an effort to make this paper self contained, we reproduce the derivation of this conjugate prior and some of its properties. 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 hierarchial 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 (2000).

The structure of the paper is as follows. In Section 2 we introduce the examples 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. Section 7 explains the underlying connection between regressions involving discrete variables and log-linear models in our Bayesian framework. In Section 8 we adapt MOSS to a clustering technique for categorical data. The relationship between MCMC and MOSS is further studied in Section 9. In Section 10 we give some concluding comments.

2. Motivating examples

We consider three examples where the data are presented under the form of a contingency table. Later we use the stochastic search algorithms developed in this paper to identify log-linear models that represent the most relevant interactions among the given variables.

Table 1. Prognostic factors for coronary heart disease (Edwards and Havranek, 1985).

				b	no		yes	
f	e	d	c	a	no	yes	no	yes
Negative	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
Positive	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

2.1. First example: the Czech autoworkers data

Table 1 contains a 2^6 table, originally analyzed by Edwards and Havranek (1985) that cross-classifies binary risk factors denoted by a, b, c, d, e, f for coronary thrombosis from a prospective epidemiological study of 1841 workers in a Czechoslovakian car factory. Here a indicates whether or not the worker "smokes", b corresponds to "strenuous mental work", c corresponds to "strenuous physical work", d corresponds to "systolic blood pressure", e corresponds to "ratio of β and α lipoproteins" and f represents "family anamnesis of coronary heart disease". This table has been extensively analyzed in the literature – see, among others, Madigan and Raftery (1994) or Dellaportas and Forster (1999).

2.2. Second example: household study in Rochdale

Our second example focuses on a cross-classification of eight binary variables relating women's economic activity and husband's unemployment from a survey of households in Rochdale – see Table 2. This study was conducted to elicit information about factors affecting the pattern of economic life and their time dynamics– see Whittaker (1990) 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 in 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.

Table 2. Women's economic activity data from Whittaker (1990). The cells counts are written in lexicographical order with h varying fastest and a varying slowest.

5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0
8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0
5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0
17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0
1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0
5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1
3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0
0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0
2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2.3. Third example: the NLTCS data

Our last example consists of a 2^{16} contingency table extracted from the "analytic" data file for National Long-Term Care Survey created by the Center of Demographic Studies at Duke University. Each dimension corresponds to a measure of disability defined by an activity of daily living, and the table contains information cross-classifying individuals aged 65 and above. The 16 dimensions of this contingency table correspond to six activities of daily living (ADLs) and ten instrumental activities of daily living (IADLs). Specifically, the ADLs are (1) eating, (2) getting in/out of bed, (3) getting around inside, (4) dressing, (5) bathing and (6) getting to the bathroom or using a toilet. The IADLs are (7) doing heavy house work, (8) doing light house work, (9) doing laundry, (10) cooking, (11) grocery shopping, (12) getting about outside, (13) travelling, (14) managing money, (15) taking medicine and (16) telephoning. For each ADL/IADL measure, subjects were classified as being either disabled (level 1) or healthy (level 0) on that measure. For a detailed description of this extract see Erosheva et al. (2008).

Dobra et al. (2003) analyze these data from a disclosure limitation perspective, while Fienberg et al. (2008) develop latent class models that are very similar to the individual-level latent mixture models from Erosheva et al. (2008). The need to consider alternatives to log-linear models for the NLTCS data comes from the severe imbalance that exists among the cell counts in this table. The largest cell count is 3853, but most of the cells (62384 or 95.19%) contain counts of zero, while 1729 (2.64%) contain counts of 1 and 1499 (0.76%) contain counts of 2. The grand total of this table is 21574, which gives a mean number of observations per cell of 0.33. This is indicative of an extremely high degree of sparsity that is characteristic of high-dimensional categorical data. For comparison, the mean number of observations per cell for the Czech Autoworkers data is 28.77, while for the Rochdale data is 2.6.

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 as defined by Dawid and Lauritzen (1993). The hyper Dirichlet 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 Madigan and Raftery (1994) or Madigan and York (1995).

When the class of possible models considered is the more general class of graphical models Markov with respect to any undirected graph or the even wider class of hierarchical models, the only priors available in the literature so far were normal priors for the log-linear parameters. Knuiman and Speed (1988) use a multivariate normal prior for the log-linear parameters. Dellaportas and Forster (1999) use a variant of this prior. King and Brooks (2001) 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. (2008) 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 (1979) (henceforth abbreviated the DY conjugate prior) for the log-linear parameters for the general class of hierarchical log-linear models.

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

3.1. Model parameterization

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 is collected in a contingency table (n) with cells indexed by combination of levels for the $|V|$ variables. We adopt the notation of Lauritzen (1996) 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 $\text{Mult}(N; (p))$ 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 to be 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 parameterization is the "baseline" or "corner" constraint parameterization 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 Agresti (1990). 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 of the cell probabilities in terms of the log-linear parameters

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

and

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

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 Darroch and Speed (1983)) 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 canonical statistics the marginal cell counts $y = (n(i_D), i_D \in \mathcal{I}_D^*, D \in \mathcal{D})$, with density, with respect to the counting measure, proportional to*

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

The proof is straightforward and is provided in the Appendix. It is important to note that

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

is the canonical parameter and

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

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

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

where $I_{\mathcal{D}}(s,\alpha) = \int_{\mathbb{R}^{d_{\mathcal{D}}}} h(\theta_{\mathcal{D}};s,\alpha)d\theta_{\mathcal{D}}$ 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

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

The corresponding hyper-parameters are:

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

From Theorem 1 of Diaconis and Ylvisaker (1979) 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 integrers.

Massam et al. (2008) study this conjugate prior in detail and give ways to elicit informative priors through the choice of hyperparameters (s,α). A non-informative conjugate prior is specified by taking all the cell entries $s(i)$ to be equal to $\frac{\alpha}{|\mathcal{I}|}$, so that

$$s(i_D) = \sum_{j\in\mathcal{I},j_D=i_D}\frac{\alpha}{|\mathcal{I}|}\,. \quad (14)$$

For the class of decomposable graphical models, this approach to constructing a conjugate prior is equivalent to eliciting hyper-Dirichlet priors – see, for example, Dawid and Lauritzen (1993) and Madigan and York (1997). While the hyper-Dirichlet priors are restricted to decomposable log-linear models, the properties of the Diaconis-Ylvisaker conjugate priors extend naturally to graphical and hierarchical log-linear models.

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

$$\widehat{\theta}_{\mathcal{D}} = \text{argmax}_{\theta_{\mathcal{D}}\in\mathbb{R}^{d_{\mathcal{D}}}} h(\theta_{\mathcal{D}};s,\alpha). \tag{15}$$

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 (15) 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 Bishop et al. (1975) and Lauritzen (1996). 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 certain 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 (Tierney and Kadane, 1986) to estimate $I_{\mathcal{D}}(s,\alpha)=\int_{\mathbb{R}^{d_{\mathcal{D}}}} h_{s,\alpha}(\theta_{\mathcal{D}})d\theta_{\mathcal{D}}$ 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

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

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

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

Let us compute the first derivative

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

Using the expression for $\frac{dp(j(G))}{d\theta(l_H)}$ derived in Massam et al. (2008), we obtain

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

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

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

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 (Dawid and Lauritzen, 1993) of the generalized hyper-Dirichlet $\pi_{\mathcal{D}}(\theta|s, \alpha)$.

Let $P_1, \ldots, P_k$ a perfect sequence of the prime components of G and let $S_2, \ldots, S_k$ be the corresponding separators, where $S_l = \left(\cup_{j=1}^{l-1} P_j \right) \cap P_l,\ \ l = 2, \ldots, k$. Dobra and Fienberg (2000) outlined fast algorithms for producing such a perfect sequence of prime components together with their separators.

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. (2008) 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 $\theta(\mathcal{D}^{S_l})$ as follows:

THEOREM 4.1. *If $\theta_{\mathcal{D}}$ follows the generalized hyper Dirichlet $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$, then the joint distribution of the parameters $\theta(\mathcal{D}^{P_l})$ and $\theta(\mathcal{D}^{S_l})$ has density*

$$\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha) = \frac{\prod_{l=2}^k I_{\mathcal{D}^{S_l}}(s^{S_l}, \alpha)}{\prod_{l=1}^k I_{\mathcal{D}^{P_l}}(s^{P_l}, \alpha)} \cdot \frac{\prod_{l=1}^k \exp\{\langle \theta(\mathcal{D}^{P_l}), s(\mathcal{D}^{P_l}) \rangle - \alpha\, k(\theta(\mathcal{D}^{P_l}))\}}{\prod_{l=2}^k \exp\{\langle \theta(\mathcal{D}^{S_l}), s(\mathcal{D}^{S_l}) \rangle - \alpha\, k(\theta(\mathcal{D}^{S_l}))\}},$$

where, for $A = S_l, R_l$, $s^A = (s(i_D), D \in \mathcal{D}^A, i_D \in \mathcal{I}_D^*)$,

$$\begin{aligned} \langle \theta(\mathcal{D}^A), s(\mathcal{D}^A) \rangle &= \sum_{D \subseteq_{\mathcal{D}^A} A} \sum_{i_D \in \mathcal{I}_D^*} \theta^A(i_D) s(i_D), \\ k(\theta(\mathcal{D}^A)) &= \log\big(1 + \sum_{D \subseteq_\ominus A} \sum_{i_D \in \mathcal{I}_D^*} \exp \sum_{F \subseteq_{\mathcal{D}^A} D} \theta^A(i_F)\big) \end{aligned}$$

and $\theta^A(i_F)$ is defined as in (2) with $p(i_H, i^*_{H^c})$ replaced by $p^A(i_H, i^*_{A \setminus H})$.

Theorem 4.1 implies that the normalizing constant $I_{\mathcal{D}}(s, \alpha)$ of $\pi_{\mathcal{D}}(\theta_{\mathcal{D}}|s, \alpha)$ is equal to

$$I_{\mathcal{D}}(s, \alpha) = \frac{\prod_{l=1}^k I_{\mathcal{D}^{P_l}}(s^{P_l}, \alpha)}{\prod_{l=2}^k I_{\mathcal{D}^{S_l}}(s^{S_l}, \alpha)}, \tag{16}$$

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 (Massam et al., 2008):

$$I_{\mathcal{D}^A}(s^A, \alpha) = \frac{\Gamma(\alpha^A_\emptyset)}{\Gamma(\alpha)} \prod_{D \in \mathcal{D}^A, i_D \in \mathcal{I}^*_D} \Gamma(\alpha^A(i_D, i^*_{A \setminus D})), \tag{17}$$

where

$$\begin{aligned} \alpha^A(i_D, i^*_{A\setminus D}) &= \sum_{A \supseteq F \supseteq D} \sum_{\substack{j_F \in \mathcal{I}^*_F \\ (j_F)_D = i_D}} (-1)^{|F \setminus D|} s(j_F), \\ \alpha^A_\emptyset &= \alpha + \sum_{D \subseteq A} (-1)^{|D|} \sum_{i \in \mathcal{I}^*_D} s(i_D). \end{aligned}$$

If A is a separator of G the subgraph G_A is always complete and we can use (17).

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 (Dobra and Fienberg, 2000). Assume that G is decomposed in 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 (16) with each $I_{\mathcal{D}^A}(s^A, \alpha)$ for $A \in \{P_1, \ldots, P_k, S_2, \ldots, S_k\}$ given by (17). 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 (Massam et al., 2008).

5. The mode oriented stochastic search (MOSS) algorithm

The Bayesian paradigm to model determination involves choosing models with high posterior probability selected from a set $\mathcal{M}$ of competing models. Godsill (2001) provides an excellent review of MCMC methods for exploring $\mathcal{M}$ such as the reversible jump sampler of Green (1995) or the product space scheme of Carlin and Chib (1995). 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 (1995) and the shotgun stochastic search (SSS) algorithm of Jones et al.

(2005) and Hans et al. (2007). If the posterior probability of a model is readily available from its marginal likelihood, up to the normalizing constant

$$\left[\sum_{m\in\mathcal{M}} \Pr(m|(n))\right]^{-1}, \tag{18}$$

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. (2005) and Hans et al. (2007) 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

$$\mathcal{M}(c) = \left\{m \in \mathcal{M} : \Pr(m|(n)) \geq c \cdot \max_{m'\in\mathcal{M}} \Pr(m'|(n))\right\}, \tag{19}$$

where $c \in (0, 1)$ and (n) is the data. We follow the Occam's window idea of Madigan and Raftery (1994) and discard models with a low posterior probability compared to the highest probability model. Raftery et al. (1997) 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 a priori equally likely, so that $\Pr(m|(n)) \propto \Pr((n)|m)$. The determination of the normalizing constant (18) 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 $\text{nbd}(m) \subset \mathcal{M}$. 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 $m_j \in \text{nbd}(m_{j-1})$ 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 terms present in the model). For details see Edwards and Havranek (1985) and Dellaportas and Forster (1999).

(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 across graphical models that involve the same number of covariates.

(c) *Decomposable log-linear models.* Here the neighborhood of a model is obtained by adding or deleting edges such that resulting graph is still decomposable – see Dawid and Lauritzen (1993) or Tarantola (2004) for details. The size 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 (18) 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 (18) 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 (1995) 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, thus $\mathcal{M}(c)$ can be exhaustively enumerated for higher values of c. We note that producing the entire set $\mathcal{M}$ is not practically possible for the examples we present 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 identiying 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, Robert (1998). 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:

$$\Pr(m_l|(n))/\left[\sum_{m' \in \mathcal{L}} \Pr(m'|(n))\right]. \tag{20}$$

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 (20) 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.

It is important to remark that 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 at 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. This feature allows MOSS to move faster than any MCMC method towards regions of high posterior probability in $\mathcal{M}$.

PROPOSITION 5.2. *MOSS(c,0,0) finds the highest posterior probability model $m_h \in \mathcal{M}$ more efficiently than any Markov chain algorithm W that moves around $\mathcal{M}$ by sampling from the posterior distribution $\{Pr(m|(n)) : m \in \mathcal{M}\}$.*

PROOF. Let us assume that MOSS and W have visited so far the same set of models $\mathcal{S}$ in $\mathcal{M}$. Let us further assume that the highest posterior probability model m_h has not been

visited yet. MOSS finds m_h at the current iteration if it selects one of its neighbours for being explored at step (b). This happens with probability

$$\left[\sum_{m\in \mathrm{nbd}(m_h)\cap\mathcal{S}} \Pr(m|(n))\right] / \left[\sum_{m'\in\mathcal{L}} \Pr(m'|(n))\right]. \tag{21}$$

Since m_h has not been reached yet, none of the neighbors of m_h has been explored and hence $\mathrm{nbd}(m_h)\cap\mathcal{S} = \mathrm{nbd}(m_h)\cap\mathcal{L}$.

At the current iteration, the Markov chain W can be in the neighborhood of m_h with probability:

$$\max\left\{\Pr(m|(n))/\left[\sum_{m'\in\mathcal{S}} \Pr(m'|(n))\right] : m\in \mathrm{nbd}(m_h)\cap\mathcal{S}\right\}. \tag{22}$$

We have $\mathcal{L}\subset\mathcal{S}$, therefore (21) is larger than (22). The probability that W finds m_h at the current iteration is actually much smaller than (22) since only one neighbor of the current model is visited and this model might not be m_h. Therefore MOSS is more likely to find m_h than W at all times.

We further study the efficiency of MOSS with respect to MCMC methods in Section 9. The examples that follow illustrate that MOSS finds the highest posterior probability models after a relatively small number of iterations.

5.1. Czech autoworkers data: revisited

We illustrate the MOSS algorithm by analyzing the Czech Autoworkers data from Table 1. MOSS was started five times from randomly generated models with $c = 0.1$, $c' = 0.001$ and a pruning probability $q = 0.1$. We use a flat conjugate prior with equal fictive cell entries as in (14). In order to assess the sensitivity of the models selected to our choice of priors, we run instances of MOSS with $\alpha \in \{1, 2, 3, 32, 64, 128\}$. This is equivalent to augmenting the actual cell counts with 1/64, 1/32, 3/64, 0.5, 1 and 2, respectively. For each value of α, we perform four separate searches as follows: (i) a search over decomposable log-linear models, (ii) a search over graphical log-linear models with marginal likelihoods estimated by decomposing the independence graph in its prime components as described in Section 4.2, (iii) a search over graphical log-linear models with marginal likelihoods estimated through a single Laplace approximation, and (iv) a search over hierarchical log-linear models. The results are shown in Tables 3 and 4. The four types of searches are labeled "Dec.", "Graph./PM", "Graph./Lapl" and "Hierar.", respectively.

We compare our results with the log-linear models selected by Dellaportas and Forster (1999) who proposed a reversible jump Markov chain Monte Carlo with normal priors for log-linear parameters and with the decomposable models selected by Madigan and Raftery (1994) who employed a hyper-Dirichlet prior for cell probabilities. For smaller values of $\alpha = 1, 2$ or 3, our most probable decomposable model $bc|ace|ade|f$ is also the best decomposable model identified by both Dellaportas and Forster (1999) and Madigan and Raftery (1994). Our most probable graphical model $ac|bc|be|ade|f$ for $\alpha = 1, 2$ or 3 in the "Graph./Lapl" search is precisely the most probable model of Dellaportas and Forster (1999) and is the second best model selected by Edwards and Havranek (1985). We remark

that both estimation methods for the marginal likelihood of graphical models yield consistent results. Our most probable hierarchical model $ac|bc|ad|ae|ce|de|f$ for $\alpha = 1, 2$ or 3 coincides with the model with the largest posterior probability identified by Dellaportas and Forster (1999). The highest probable models selected by us and by Dellaportas and Forster (1999) are extremely consistent for small choices of α.

Increasing α leads to the inclusion of higher order interaction terms in the corresponding log-linear models. This implies that α effectively penalizes for an increased model complexity. Sparser log-linear models are identified by decreasing α and these are precisely the models we are looking for in higher-dimensional contingency tables having most counts equal to zero. Such tables cannot support more complex interactions due to the small number of observed samples. For a fixed value of α, the highest probable log-linear models also become sparser as we sequentially relax the structural constraints from decomposable to graphical and hierarchical. Tables 3 and 4 show that the most probable graphical (hierarchical) models can be obtained by dropping some of the second order interaction terms in the most probable decomposable (graphical) models.

Massam et al. (2008) present the highest probable log-linear models identified by employing MC^3 instead of MOSS. For each value of α and each class of log-linear models, they run four separate Markov chains from random starting models for $25,000$ iterations with a burn-in of $5,000$ iterations. The top models identified by MC^3 are precisely the top models identified by MOSS. The posterior probabilities of each model determined in the MC^3 search are slightly smaller than the posterior model probabilities as reported by MOSS since lower posterior probability models that fall outside $\mathcal{M}(0.1)$ are discarded. Eliminating these models has no consequence on the top models identified. However, the highest posterior probability models are more likely to coincide with the median log-linear model with respect to $\mathcal{M}(c)$. A median log-linear model contains those interaction terms having a posterior inclusion probability greater than 0.5. Such models are hierarchical irrespective of the class of log-linear models considered. The median log-linear models associated with the entire set of candidate models $\mathcal{M}$ might contain spurious interaction terms that do not appear in the highest probable models. This remark is even more important in the case of hierarchical log-linear models whose individual posterior probability tend to be small because only one or at most two interactions terms differentiate models having close posterior probabilities.

Table 5 gives the number of models visited by MOSS before its completion. In most cases only a couple of hundred models need to be evaluated to identify the highest probability models. The number of models visited by MC^3 in each search was $30,000$. While it is extremely likely that a much smaller number of MCMC iterations were needed to get to the top models, actually determining when convergence has been reached is traditionally a tedious process. MOSS provides a simple way to avoid the need to evaluate convergence because it has an implicit stopping rule. This is another reason why MOSS is more efficient than MC^3 and this efficiency becomes even more important in the analysis of higher-dimensional datasets. Table 5 also shows that the number of models evaluated increases as α increases since the highest posterior probability models contain more complex interaction terms.

A very interesting question is whether there exists evidence between the family history of coronary heart disease (variable f) and the five risk factors a, b, c, d and e. Whittaker (1990) page 263 chooses the graphical model $abce|ade|bf$ that links f with b – strenuous mental work. The most probable models identified by Dellaportas and Forster (1999), Madigan and Raftery (1994) or Edwards and Havranek (1985) indicate the independence of f from the other risk factors. Their findings are consistent with the models we identify

Table 3. The models with the highest posterior probabilities identified by MOSS for the Czech autoworkers data when $\alpha \in \{1, 2, 3\}$. We give the models whose normalized posterior probabilities are greater than 0.05. The median log-linear models are labeled "med."

Search	$\alpha = 1$		$\alpha = 2$		$\alpha = 3$	
Dec.	$bc\|ace\|ade\|f$	0.370	$bc\|ace\|ade\|f$	0.342	$bc\|ace\|ade\|f$	0.425
	$bc\|ace\|de\|f$	0.155	$bc\|ace\|de\|f$	0.231	$bc\|ace\|ade\|bf$	0.211
	$bc\|ad\|ace\|f$	0.151	$bc\|ace\|de\|bf$	0.125	$bc\|ace\|de\|f$	0.145
	$ac\|bc\|be\|de\|f$	0.089	$bc\|ad\|ace\|f$	0.094	$bc\|ace\|ade\|ef$	0.089
	$bc\|ace\|de\|bf$	0.076	$bc\|ace\|de\|bf$	0.085	$bc\|ace\|de\|bf$	0.072
	$ac\|bc\|ae\|de\|f$	0.068	$bc\|ace\|ade\|ef$	0.053	$bc\|ad\|ace\|f$	0.059
	$bc\|ace\|de\|f$	med.	$bc\|ace\|ade\|f$	med.	$bc\|ace\|ade\|f$	med.
Graph./PM	$ac\|bc\|ae\|be\|de\|f$	0.577	$ac\|bc\|ae\|be\|de\|f$	0.482	$ac\|bc\|ae\|be\|de\|f$	0.432
	$ac\|bc\|ad\|ae\|be\|f$	0.235	$ac\|bc\|ad\|ae\|be\|f$	0.196	$ac\|bc\|ae\|be\|de\|bf$	0.215
	$ac\|bc\|ae\|be\|de\|bf$	0.119	$ac\|bc\|ae\|be\|de\|bf$	0.176	$ac\|bc\|ad\|ae\|be\|f$	0.176
	$ac\|bc\|d\|ae\|be\|f$	0.070	$ac\|bc\|ae\|be\|de\|ef$	0.074	$ac\|bc\|ae\|be\|de\|ef$	0.090
			$ac\|bc\|ad\|ae\|be\|bf$	0.072	$ac\|bc\|ad\|ae\|be\|bf$	0.087
	$ac\|bc\|ae\|be\|de\|f$	med.	$ac\|bc\|ae\|be\|de\|f$	med.	$ac\|bc\|ae\|be\|de\|f$	med.
Graph./Lapl	$ac\|bc\|be\|ade\|f$	0.391	$ac\|bc\|be\|ade\|f$	0.454	$ac\|bc\|be\|ade\|f$	0.485
	$ac\|bc\|ae\|be\|de\|f$	0.264	$ac\|bc\|be\|ade\|bf$	0.187	$ac\|bc\|be\|ade\|bf$	0.245
	$ac\|bc\|be\|ade\|bf$	0.114	$ac\|bc\|ae\|be\|de\|f$	0.154	$ac\|bc\|ae\|be\|de\|f$	0.111
	$ac\|bc\|ad\|ae\|be\|f$	0.108	$ac\|bc\|be\|ade\|ef$	0.079	$ac\|bc\|be\|ade\|ef$	0.103
	$ac\|bc\|ae\|be\|de\|bf$	0.077	$ac\|bc\|ae\|be\|de\|bf$	0.064	$ac\|bc\|ae\|be\|de\|bf$	0.056
			$ac\|bc\|ad\|ae\|be\|f$	0.063		
	$ac\|bc\|be\|ade\|f$	med.	$ac\|bc\|be\|ade\|f$	med.	$ac\|bc\|be\|ade\|f$	med.
Hierar.	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.392	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.298	$ac\|bc\|ad\|ae\|ce\|de\|f$	0.256
	$ac\|bc\|ad\|ae\|be\|de\|f$	0.246	$ac\|bc\|ad\|ae\|be\|de\|f$	0.187	$ac\|bc\|ad\|ae\|be\|de\|f$	0.161
	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	0.124	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	0.133	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	0.140
	$ac\|bc\|ad\|ae\|ce\|de\|bf$	0.114	$ac\|bc\|ad\|ae\|ce\|de\|bf$	0.123	$ac\|bc\|ad\|ae\|ce\|de\|bf$	0.129
	$ac\|bc\|ad\|ae\|be\|de\|bf$	0.071	$ac\|bc\|ad\|ae\|be\|de\|bf$	0.077	$ac\|bc\|ad\|ae\|be\|de\|bf$	0.081
			$ac\|bc\|ad\|ae\|be\|ce\|de\|bf$	0.055	$ac\|bc\|ad\|ae\|be\|ce\|de\|bf$	0.071
			$ac\|bc\|ad\|ae\|ce\|de\|ef$	0.052	$ac\|bc\|ad\|ae\|ce\|de\|ef$	0.055
	$ac\|bc\|ad\|ae\|ce\|de\|f$	med.	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	med.	$ac\|bc\|ad\|ae\|be\|ce\|de\|f$	med.

Table 4. The models with the highest posterior probabilities identified by MOSS for the Czech Autoworkers data when $\alpha \in \{32, 64, 128\}$. We give the models whose normalized posterior probabilities are greater than 0.05.

Search	$\alpha = 32$		$\alpha = 64$		$\alpha = 128$	
Dec.	$bc\|ace\|ade\|bf$	0.169	$ace\|bce\|ade\|bcf$	0.134	$ace\|bce\|ade\|bcf$	0.359
	$ace\|bce\|ade\|bf$	0.123	$ace\|bce\|ade\|bf$	0.118	$ace\|ade\|bcf\|cef$	0.133
	$bc\|ace\|ade\|f$	0.077	$ace\|ade\|bcf$	0.081	$abc\|ace\|ade\|bcf$	0.105
	$abc\|ace\|ade\|bf$	0.075	$bc\|ace\|ade\|bf$	0.071	$abc\|abe\|ade\|bcf$	0.104
	$bc\|ace\|ade\|ef$	0.071	$abc\|ace\|ade\|acf$	0.062	$ace\|ade\|acf$	0.089
	$abc\|abe\|ade\|bf$	0.057	$abc\|ace\|ade\|bf$	0.055	$abce\|ade\|acf$	0.060
	$ace\|bce\|ade\|f$	0.056	$abc\|abe\|ade\|acf$	0.052	$ace\|ade\|bcef$	0.051
	$ace\|bce\|ade\|ef$	0.051			$ace\|ade\|abcf$	0.050
	$bc\|ace\|ade\|bf$	med.	$bc\|be\|ace\|ade\|bf$	med.	$be\|ace\|ade\|bcf$	med.
Graph./PM	$bc\|ace\|ade\|bf$	0.093	$ace\|bce\|ade\|bcf$	0.091	$ace\|bce\|ade\|bcf$	0.280
	$ac\|bc\|ade\|bde\|bf$	0.069	$ace\|bce\|ade\|bf$	0.080	$ace\|bce\|ade\|bde\|bcf$	0.138
	$ace\|bce\|ade\|bf$	0.068	$ace\|ade\|acf$	0.055	$ace\|ade\|bcf\|cef$	0.104
	$bc\|acd\|be\|ade\|bf$	0.058			$abc\|ace\|ade\|bcf$	0.082
	$bc\|bd\|ace\|ade\|bf$	0.053			$abc\|abe\|ade\|bcf$	0.081
					$ace\|ade\|bcf$	0.070
	$bc\|be\|ace\|ade\|bf$	med.	$bc\|be\|ace\|ade\|bf$	med.	$be\|ace\|ade\|bcf$	med.
Graph./Lapl	$ac\|bc\|be\|ade\|bf$	0.303	$ac\|bc\|be\|ade\|bf$	0.162	$ac\|be\|ade\|bcf$	0.161
	$ac\|bc\|be\|ade\|f$	0.167	$ac\|be\|ade\|bcf$	0.128	$ace\|bce\|ade\|bcf$	0.114
	$ac\|bc\|be\|ade\|ef$	0.127	$ac\|bc\|be\|ade\|af\|bf$	0.068	$ac\|be\|ade\|bcf\|df$	0.109
	$ac\|bc\|be\|ade\|af\|bf$	0.091	$ac\|bc\|be\|ade\|bf\|df$	0.068	$ace\|bce\|ade\|bcf\|df$	0.077
	$ac\|bc\|be\|ade\|bf\|df$	0.084	$ac\|bc\|be\|ade\|ef$	0.068	$ac\|ade\|bcf\|bef$	0.069
	$ac\|be\|ade\|acf$	0.059	$ac\|bc\|be\|ade\|f$	0.057	$ac\|ade\|bcf\|bef\|def$	0.064
			$ac\|be\|ade\|bcf\|df$	0.054		
	$ac\|bc\|be\|ade\|bf$	med.	$ac\|bc\|be\|ade\|bf$	med.	$ac\|be\|ade\|bcf$	med.
Hierar.	$ac\|bc\|ad\|ae\|be\|ce\|de\|bf$	0.071	$ac\|bc\|ad\|ae\|be\|ce\|de\|bf$	0.023	$ac\|ad\|ae\|be\|ce\|de\|bcf\|ef$	0.012
	$ac\|bc\|ad\|ae\|be\|ce\|de\|bf$	med.	$ac\|bc\|ad\|ae\|be\|ce\|de\|bf$	med.	$ac\|ad\|ae\|be\|ce\|de\|bcf\|df\|ef$	med.

Table 5. The minimum, median and maximum number of models evaluated by MOSS for the Czech Autoworkers data for non-informative priors induced by $\alpha \in \{1, 2, 3, 32, 64, 128\}$ across five search replicates.

	α					
Search	1	2	3	32	64	128
Dec.	85|177|397	179|216|454	117|236|340	256|349|416	374|381|415	201|255|294
Graph./PM	95|167|223	267|343|442	191|201|394	718|1029|1266	621|1048|1139	259|452|721
Graph./Lapl	217|311|637	209|237|478	220|315|439	195|365|788	652|743|806	420|621|859
Hierar.	636|752|834	701|744|1045	548|811|877	1446|1544|1767	3417|3954|4072	6296|6372|6808

Table 6. Posterior inclusion probabilities for the edge bf for various choice of priors and classes of log-linear models as determined by MOSS for the Czech autoworkers data.

	α					
Search	1	2	3	32	64	128
Dec.	0.076	0.244	0.283	0.522	0.715	1
Graph./PM	0.119	0.248	0.302	0.533	0.697	1
Graph./Lapl	0.190	0.251	0.301	0.616	0.785	1
Hierar.	0.186	0.263	0.290	0.749	0.912	1

for smaller values of α. However, as we increase the grand total α in the prior fictive table we employ, a direct link between b and f appears in our highest probable models. Table 6 shows the posterior inclusion probability of the first order interaction between b and f for various choices of α and structural model constraints. Table 6 seems to confirm Whittaker's findings as the posterior probability of the edge bf increases from 0.076 to almost 1. This edge does not appear in sparser models corresponding with smaller values of α because there are stronger associations among the five risk factors than between a particular risk factor and the family history of coronary heart disease. The first-order interaction bf enters the top models only if the penalty for model complexity is decreased.

5.2. Household study in Rochdale: revisited

Next we use MOSS to analyze the Rochdale data presented in Table 2. Whittaker (1990) 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 the all two-way interaction model and sequentially eliminates edges based on their deviances. All the higher-order interactions were discarded up front. Whittaker chooses the model

$$fg|ef|dh|dg|cg|cf|ce|bh|be|bd|ag|ae|ad|ac. \tag{23}$$

To the extent of the authors' knowledge, there was no other 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 models search was started from the top decomposable models identified by MOSS, while the hierarchical models 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. We have already seen for the Czech autoworkers data that there is a strong relationship among the highest posterior probability models associated with nested classes of log-linear models.

Table 7 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 7 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 extremely efficient since explicit formulas exist in this case, hence this relatively modest number of visited models gives a good indication of the performance of MOSS. Since numerical approximations to marginal likelihoods have to be used in the graphical and hierarchical case, it is imperative to attempt to reduce the number of models that are visited due to 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 (Dellaportas and Forster, 1999) by visiting less than $2,000$ 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 computational 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 non-linear transformations of $\theta_{\mathcal{D}}$.

To this end, Gelman et al. (2004) and Schafer (1997) 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 extremely 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 with certain shape parameters. Piccioni (2000) 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 (2007) 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 Asci and Piccioni (2007). 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 $\theta^{(0)} \in \mathbb{R}^{|\mathcal{E}|} = \mathbb{R}^{2^{|V|}}$

Table 7. 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.	

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

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

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

$$\theta^{A_l}(i_E) = \log \prod_{F \subseteq E} p^{A_l}(i_F, i^*_{A_l \setminus F})^{(-1)^{|E \setminus F|}}. \tag{24}$$

(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

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

- 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 illustrate the use of Bayesian IPF to generate 5,000 random draws from the conjugate posterior and the conjugate prior distribution associated with the log-linear model $ac|bc|ad|ae|ce|de|f$ – see Figure 1. From Table 3 we learned that this was the top hierarchical log-linear model for the Czech autoworkers data for $\alpha = 1$. There are twelve independent parameters: $\theta(a)$, $\theta(b)$, $\theta(c)$, $\theta(ac)$, $\theta(bc)$, $\theta(d)$, $\theta(ad)$, $\theta(e)$, $\theta(ae)$, $\theta(ce)$, $\theta(de)$ and $\theta(f)$. This order was used to number the θ's from 1 to 12 in Figure 1. Estimates of the posterior and prior density of each parameter are plotted in blue, while the dotted black

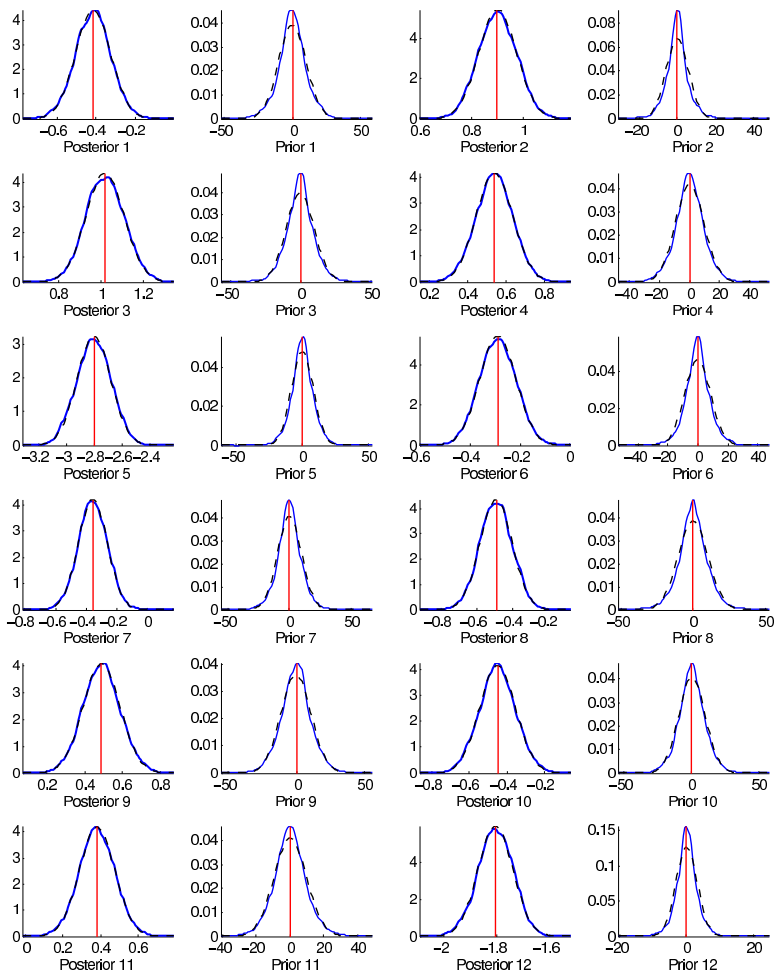

Fig. 1. Posterior and prior density estimates (solid blue lines) for the twelve free parameters of the log-linear model $ac|bc|ad|ae|ce|de|f$ for the Czech Autoworkers data. The dotted black lines give the sample normal approximation, while the red lines represent the mode of these distributions as estimated using IPF.

lines represent the corresponding sample normal approximations. The IPF algorithm was used to identify the mode of the joint posterior and its conjugate prior. The mode estimate of each parameter is represented with a red line. Remark that the posterior densities of the θ's are very close to their normal approximations and that IPF correctly identifies the posterior modes. The priors are always centered at zero with a high variance, hence they are proper and non-informative as we would expect. They also tend to have slightly heavier tails than their normal approximations.

7. Regressions induced by log-linear models

We consider the problem of studying how a subset of response variables X_A, $A \subset V$ are influenced by the remaining covariates X_{A^c} where $A^c = V \setminus A$. In particular, we are interested in transforming a log-linear model into the regression of X_A on X_{A^c}. Since a log-linear model gives a parsimonious representation of the joint distribution of all variables in a contingency table, the dependencies that might exist among the explanatory variables are taken into account in the implied conditional $[X_A|X_{A^c}]$. Moreover, it is likely that the log-linear interaction terms involving one or more response variables X_v, $v \in A$, might contain only a subset of the explanatory variables X_E, $E \subset A^c$. This means that a variable selection step is performed and the full regression $[X_A|X_{A^c}]$ reduces to $[X_A|X_E]$. Similar ideas have been previously discussed by Agresti (1990) who describes the relationship between log-linear and logit models.

We want to compute

$$\log p(i_A|i_{A^c}) = \log p(i_A, i_{A^c}) - \log \sum_{j_A \in \mathcal{I}_A} p(j_A, i_{A^c}) \, . \tag{25}$$

We assume that $(i_A, i_{A^c}) \neq i^*$. There exists $B \subset V$ such that $(i_A, i_{A^c}) = (i_B, i^*_{B^c})$ such that $i_B \in \mathcal{I}^*_B$. With the usual notation $(i_B, i^*_{B^c}) = i(B)$, we have

$$p(i_A, i_{A^c}) = p(i(B)) = \frac{\exp \sum_{F \subseteq_{\mathcal{D}} B} \theta(i_F)}{1 + \sum_{E \in \mathcal{E}_\ominus} \sum_{i_E \in \mathcal{I}^*_E} \exp \sum_{F \subseteq_{\mathcal{D}} E} \theta(i_F)}. \tag{26}$$

Similarly, each $(j_A, i_{A^c}) = i(H_{j_A})$ for some $H_{j_A} \subseteq V$ such that $H_{j_A} \cap A^c = B \cap A^c$ and

$$p(j_A, i_{A^c}) = p(i(H_{j_A})) = \frac{\exp \sum_{F \subseteq_{\mathcal{D}} H_{j_A}} \theta((j_A, i_{A^c})_F)}{1 + \sum_{E \in \mathcal{E}_\ominus} \sum_{i_E \in \mathcal{I}^*_E} \exp \sum_{F \subseteq_{\mathcal{D}} E} \theta(i_F)}. \tag{27}$$

From (25), (26) and (27), we obtain

$$\log p(i_A|i_{A^c}) = \sum_{F \subseteq_{\mathcal{D}} B} \theta(i_F) - \log \sum_{j_A \in \mathcal{I}_A} \exp \sum_{F \subseteq_{\mathcal{D}} H_{j_A}} \theta((j_A, i_{A^c})_F), \tag{28}$$

that is the desired formula for the regression of X_A on X_{A^c}. In the case of binary data, (28) becomes

$$\log p(i_A|i_{A^c}) = \sum_{F \subseteq_{\mathcal{D}} B} \theta(F) - \log \sum_{G \subseteq A} \exp \sum_{F \subseteq_{\mathcal{D}} G \cup (A^c \cap B)} \theta(F). \tag{29}$$

In the case where there is only one response variable X_γ, i.e., $A = \{\gamma\}$ for some $\gamma \in V$, instead of considering (29), we may consider the odds ratio which will then be equal to

$$\log\left(\frac{p(i_\gamma|i_{B\setminus\gamma}, i^*_{B^c})}{p((i^*)_\gamma|i_{B\setminus\gamma}, i^*_{B^c})}\right) = \frac{\sum_{D\subseteq_{\mathcal{D}} B}\theta(D)}{\sum_{D\subseteq_{\mathcal{D}} B\setminus\{\gamma\}}\theta(D)} = \theta_\gamma + \sum_{(\gamma\cup D)\subseteq_{\mathcal{D}} B}\theta_{\gamma\cup D};. \tag{30}$$

Example. We exemplify these results with the problem of predicting wife's economic activity a in the Rochdale data. Whittaker (1990) page 285 considers the log-linear model $ac|ad|ae|ag$ induced by the generators of (23) 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:

$$\log\frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} = \text{const.} - 1.33c - 1.32d + 0.69e - 2.17g. \tag{31}$$

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 7)

$$fg|ef|dg|cg|cf|ce|be|bdh|ag|ae|ad|ac \tag{32}$$

are again ac, ad, ae and ag which, according to (30), yield the regression equation

$$\log\frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} = \theta(a) + \theta(ac) + \theta(ad) + \theta(ae) + \theta(ag). \tag{33}$$

Using Bayesian IPF to produce $10,000$ draws from the posterior probability associated with the log-linear model (32), we estimate the regression equation (33) 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 (31), 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 (32), while Whittaker used the five-way marginal associated with a, c, d, e and g to fit the log-linear model $ac|ad|ae|ag$.

8. Clustering discrete data with MOSS

MOSS seems to scale well and facilitate an efficient determination of hierarchical log-linear models for dichotomous eight-way tables. Unfortunately the number of candidate models increases way too fast with the inclusion of only a few additional categorical variables to allow MOSS to perform equally well for higher-dimensional tables such as the NLTCS data presented in Section 2.3. Model selection is further compounded by the extremely small number of observed samples that makes most of the cells to contain zero counts. Recall that 95.19% of the counts in the NLTCS data were zero. To address these issues we develop a clustering algorithm that breaks the full table into marginals involving non-overlapping subsets of variables. These smaller dimensional tables are less sparse and can be analyzed separately. Hu and Johnson (2007) present an MCMC approach to identify log-linear models

Table 8. The cluster log-linear models with the highest posterior probabilities identified by MOSS for the Czech autoworkers data. The non-informative conjugate priors are induced by $\alpha \in \{1, 2, 3, 32, 64, 128\}$. We report the models whose normalized posterior probabilities are greater than 0.05.

$\alpha = 1$		$\alpha = 2$		$\alpha = 3$		$\alpha = 32$		$\alpha = 64$		$\alpha = 128$	
$abc\|de\|f$	0.458	$abc\|de\|f$	0.519	$abc\|de\|f$	0.530	$abce\|d\|f$	0.738	$abce\|d\|f$	0.654	$abce\|d\|f$	0.378
$bc\|ade\|f$	0.368	$bc\|ade\|f$	0.419	$bc\|ade\|f$	0.429	$abce\|df$	0.256	$abce\|df$	0.345	$abce\|df$	0.320
$bc\|d\|ae\|f$	0.107									$abcef\|d$	0.254
$abc\|d\|e\|f$	0.055										
$abc\|de\|f$	med.	$abc\|de\|f$	med.	$abc\|de\|f$	med.	$abce\|d\|f$	med.	$abce\|d\|f$	med.	$abce\|d\|f$	med.

having non-overlapping minimal sufficient statistics.

We are interested in log-linear models m whose generators $\{P_1, P_2, \ldots, P_k\}$, $k \geq 1$, represent a clustering of the variables in V, i.e. $P_1 \cup \ldots \cup P_k = V$ and $P_i \cap P_k = \emptyset$ for $i \neq j$. Such a log-linear model is clearly decomposable. Its cliques are precisely $P_1, P_2, \ldots, P_k$, while its separators are $k - 1$ empty sets. The marginal likelihood of m is easily calculated with the formulas from Section 4.3. We call m a *cluster* log-linear model. MOSS can be used to perform a stochastic search over the class of cluster log-linear models. We follow the ideas in Hu and Johnson (2007) and define the neighborhood of the model m to be generated through two types of moves:

(a) *Split move*: Replace a cluster P_j with two sub-clusters $P_j^{(1)}$ and $P_j^{(2)}$ such that $P_j^{(1)} \cup P_j^{(2)} = P_j$ and $P_j^{(1)} \cap P_j^{(2)} = \emptyset$.
(b) *Merge move*: Replace two clusters P_i and P_j, $i \neq j$, with the cluster $P_i \cup P_j$.

To avoid clusters with too many variables, we allow a merge move only if the size of the resulting cluster is below a certain cutoff q'. Visiting a cluster model m involves producing all the neighbors of m, that is, splitting each cluster in every possible way and attempting to merge any two clusters.

Examples. We use MOSS to cluster the variables in the three datasets from Section 2. We run five instances of MOSS starting from random clusters with $c = 0.333$, $c' = 0.001$ and a pruning probability of $q = 0.1$.

Table 8 gives the top cluster models for the Czech Autoworkers data. We allowed the clusters to have a maximum size $q' = 6$. Most of these clusters also appear as interaction terms in the highest probable decomposable, graphical or log-linear models from Tables 3 and 4. For the Rochdale data we took $q' = 8$ and determined the cluster model $acg|bdh|ef$ with a posterior probability of almost 1. All the terms in this model also appear in the top models from Table 7.

Next we proceed to the sixteen-dimensional NLTCS data. A choice of $q' = 16$ leads to a model with only one cluster containing all sixteen variables. For computational reasons, we set $q' = 8$ and identify the cluster model

$$1, 5, 11, 12, 13, 14, 15, 16|2, 3, 4, 6, 7, 8, 9, 10$$

with a normalized posterior probability equal to one. This clustering seems to make sense since the first group contains activities that relate a person to the outside world, while the second group contains activities that take place exclusively around the house. The "outdoor" cluster contains two ADLs and six IADLs, while the "indoor" cluster is more

balanced having four ADLs and four IADLs. The two marginal tables associated with these clusters are less sparse since their mean number of observations per cell is 84.3. The "outdoor" marginal has 213 non-zero counts with a largest count of 6860 in the $(0, 0, \ldots, 0)$ cell. The "indoor" marginal has 182 non-zero counts with a largest count of 5268 in the $(0, 0, \ldots, 0)$ cell. This means that log-linear models are likely to be suitable for representing associations in these two eight-way tables.

We run MOSS from five starting points with $c = 0.333$, $c' = 0.001$ and a pruning probability of 0.01. The top hierarchical model for the "outdoor" marginal has a normalized posterior probability of 0.754:

$$\begin{aligned} &13, 15, 16|13, 14|12, 16|12, 13|11, 15, 16|11, 13, 15|11, 12, 14, 15| \\ &\quad 5, 14, 16|5, 14, 15|5, 12, 14|5, 11, 15|1, 16|1, 13|1, 12|1, 5. \end{aligned} \tag{34}$$

The top hierarchical model for the "indoor" marginal has a normalized posterior probability of 0.358 and coincides with the median hierarchical model in $\mathcal{M}(0.333)$:

$$\begin{aligned} &8, 9, 10|7, 8, 10|7, 8, 9|6, 9|6, 7, 10|4, 7|4, 6, 8, 9| \\ &\quad 3, 7|3, 4, 8|3, 4, 6|2, 9|2, 8|2, 7|2, 4, 10|2, 3. \end{aligned} \tag{35}$$

Many of the interactions present in these two log-linear models seem to be reasonable. For example, the two nested IADLs 7 and 8 (doing heavy and light housework) belong to the same cluster and appear together in two second-order terms in (35). The term "nested" refers to the fact that an individual who is incapable of doing light housework is also incapable of doing heavy houwsework.

We assess the fit of these models by drawing 10000 samples from the posterior distributions of their parameters using Bayesian IPF. The p-value for model (34) is 0 based on a χ^2 value of 587.57 on 212 degrees of freedom, while the p-value for model (35) is 0 based on a χ^2 value of 414.62 on 212 degrees of freedom.

By joining the interactions in (34) and (35) we obtain a log-linear model for the complete 2^{16} table. However, this log-linear model does not have a good fit. The posterior mean of the cell probability $(0, 0, \ldots, 0)$ in the "outdoor" marginal is 0.242. The corresponding value in the "indoor" marginal is 0.313. This leads to a fitted value for the $(0, 0, \ldots, 0)$ cell of the NLTCS data of 1631.76 that is not close to the observed count of 3853.

Breaking the initial table in several non-overlapping marginals might offer some insight about the underlying associations that exist, but some other relevant associations could be lost when the clusters are created. This is precisely why the unrestricted cluster search returned the complete 2^{16} table.

9. MCMC vs. MOSS

We would like to gain further insight about the relative efficiency of MOSS with respect to MCMC stochastic search algorithms by studying how much of the total posterior probability is actually covered by the subset $\mathcal{M}(c)$, $c \in (0, 1)$ of models with the highest posterior probability. That is, we want to determine the ratio:

$$\left[\sum_{m \in \mathcal{M}(c)} \Pr(m|(n)) \right] / \left[\sum_{m' \in \mathcal{M}} \Pr(m'|(n)) \right]. \tag{36}$$

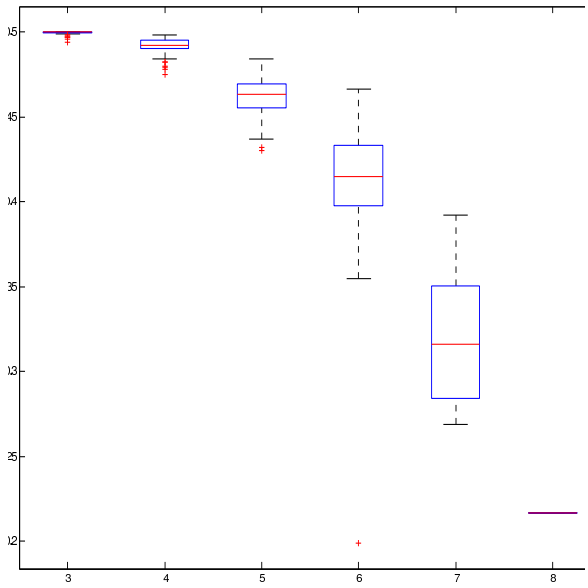

Fig. 2. Ratios between the posterior probability of the models in $\mathcal{M}(0.1)$ and the total probability mass over the hierarchical models of 219 marginal tables derived from the Rochdale data. The ratios (y-axis) are grouped by the dimension of the marginals (x-axis).

In order to evaluate this ratio we use the MC^3 algorithm of Madigan and York (1995) which can be briefly described as follows. This algorithm constructs an irreducible Markov chain m_t, $t = 1, 2, \ldots$ with state space $\mathcal{M}$ and equilibrium distribution $\{\Pr(m|(n)) : m \in \mathcal{M}\}$. If the chain is in state m_t at time t, a candidate model m' is drawn from a uniform distribution on $\text{nbd}(m_t)$. The chain moves in state m' at time $t+1$, i.e. $m_{t+1} = m'$ with probability

$$\min\left\{1, \frac{\Pr((n)|m_{t+1})/\#\text{nbd}(m_{t+1})}{\Pr((n)|m_t)/\#\text{nbd}(m_t)}\right\}, \tag{37}$$

where $\#\text{nbd}(m)$ denotes the number of neighbors of m. Otherwise the chain does not move, i.e. we set $m_{t+1} = m_t$. In (37) it was assumed that all models are apriori equally likely. The number of times the chain hits a model in $\mathcal{M}(c)$ divided by the total number of iterations represents an estimate of (36).

We used MOSS to determine the top hierarchical log-linear models in $\mathcal{M}(c)$ with $c = 0.1$ for 219 marginal tables derived from the Rochdale data: 56 three-way tables, 70 four-way tables, 56 five-way tables, 28 six-way tables, 8 seven-way tables and one eight-way table. For each of these marginal tables, we run MC^3 from four different starting points for $10,000$ iterations with a burn-in of $2,500$ iterations. Figure 2 shows the resulting estimates of (36) grouped by the dimension of the tables analyzed. For three-way tables the coverage of probability space seems to be around 0.5, but it decreases to about 0.4 for six-way tables, 0.3 for seven-way tables and to approximately 0.2 for the full eight-way table. It seems appropriate to infer that the ratio (36) goes to zero as the dimension of the contingency tables increases while the sample size remains fixed.

MCMC algorithms identify models with high posterior probability by sampling from the posterior distribution over the space of candidate models. If there are many models with low

posterior probability with respect to the model with the highest posterior probability, and their total posterior probability dominates the space then most of the MCMC iterations are spent visiting such models. This makes any Markov chain likely to be extremely inefficient for high-dimensional datasets because sampling from $\{\Pr(m|(n)) : m \in \mathcal{M}\}$ becomes a different question than the efficient determination of $\mathcal{M}(c)$. MOSS is designed to solve the latter problem, while MCMC algorithms solve the former problem.

10. Conclusions

In this paper we showed that the combination of MOSS and the conjugate prior for log-linear parameters of Massam et al. (2008) 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. We attempted to use MC^3 to find hierarchical log-linear models for the Rochdale data, but we did not manage to obtain results worth mentioning. 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. The clustering technique we proposed is able to quickly identify the most relevant groups of variables and scales to sparse datasets involving a larger number of discrete variables.

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/`

The current implementation of MOSS is written only for dichotomous contingency tables. The methods presented in this paper hold for arbitrary multi-way cross-classifications and our code can therefore be extended to polychotomous data in a straighforward albeit time consumming manner.

References

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

Asci, C. and M. Piccioni (2007). Functionally compatible local characteristics for the local specification of priors in graphical models. *Scand. J. Statist. 34*, 829–840.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. M.I.T. Press. Cambridge, MA.

Carlin, B. P. and S. Chib (1995). Bayesian model choice via Markov chain Monte Carlo. *J. R. Stat. Soc. Ser. B 57*, 473–484.

Clyde, M. and E. I. George (2004). Model uncertainty. *Statist. Sci. 19*, 81–94.

Darroch, J. and T. Speed (1983). Additive and multiplicative models and interaction. *Ann. Statist. 11*, 724–738.

Dawid, A. P. and S. L. Lauritzen (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist. 21*, 1272–1317.

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

Dellaportas, P. and C. Tarantola (2005). Model determination for categorical data with factor level merging. *J. R. Stat. Soc. Ser. B 67*, 269–283.

Diaconis, P. and D. Ylvisaker (1979). Conjugate priors for exponential families. *Ann. Statist. 7*, 269–281.

Dobra, A., E. A. Erosheva, and S. E. Fienberg (2003). Disclosure limitation methods based on bounds for large contingency tables with application to disability data. In e. H. Bozdogan (Ed.), *Proceedings of Conference on the New Frontiers of Statistical Data Mining*, pp. 93–116. CRC Press.

Dobra, A. and S. E. Fienberg (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Natl. Acad. Sci. 97*, 11185–11192.

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

Erosheva, E. A., S. E. Fienberg, and C. Joutard (2008). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat. 1*, 502–537.

Fienberg, S. E., P. Hersh, A. Rinaldo, and Y. Zhou (2008). Maximum likelihood estimation in latent class models for contingency table data. In P. Gibilisco, E. Riccomagno, M. P. Rogantin, and e. Wynn, H. P. (Eds.), *Algebraic and Geometric Methods in Statistics*. Cambridge University Press. forthcoming.

Fienberg, S. E. and A. Rinaldo (2007). Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference 137*, 3430–3445.

Gelman, A., J. B. Carlin, H. S. Stern, and D. Rubin (2004). *Bayesian Data Analysis* (Second ed.). Texts in Statistical Science Series. Chapman & Hall.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comput. Graph. Statist. 10*, 1–19.

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

Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc. 102*, 507–516.

Hu, J. and V. E. Johnson (2007, October). Log-linear models for gene association. Technical Report 38, UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series. URL: http://www.bepress.com/mdandersonbiostat/paper38.

Jones, B., C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci. 20*, 388–400.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *J. Amer. Statist. Assoc. 90*, 773–795.

King, R. and S. P. Brooks (2001). Prior induction for log-linear models for general contingency table analysis. *Ann. Statist. 29*, 715–747.

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

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Madigan, D. and A. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc. 89*, 1535–1546.

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

Madigan, D. and J. York (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika 84*, 19–31.

Massam, H., J. Liu, and A. Dobra (2008). A conjugate prior for discrete hierarchical log-linear models. Available from http://arxiv.org/abs/0711.1609.

Piccioni, M. (2000). Independence structure of natural conjugate densities to exponential families and the Gibbs sampler. *Scand. J. Statist. 27*, 111–127.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc. 92*, 179–191.

Robert, C. P. (1998). *Discretization and MCMC Convergence Assessment*, Volume 135 of *Lecture Notes in Statistics*. Springer-Verlag.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Tarantola, C. (2004). MCMC model determination for discrete graphical models. *Statistical Modelling 4*, 39–61.

Tierney, L. and J. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc. 81*, 82–86.

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

11. Appendix

Proof of Lemma 3.2

We have

$$\begin{aligned}
\prod_{i\in\mathcal{I}} p(i)^{n(i)} &= p_\emptyset^{n(i^*)} \prod_{E\in\mathcal{E}_\ominus} \prod_{i_E\in\mathcal{I}_E^*} p(i(E))^{n(i(E))}, \\
&= p_\emptyset^{n(i^*)} \prod_{E\in\mathcal{E}_\ominus} \prod_{i_E\in\mathcal{I}_E^*} \Big(\exp\sum_{F\subseteq E}\theta(i_F)\Big)^{n(i(E))}, \\
&= \prod_{E\in\mathcal{E}_\ominus} \prod_{i_E\in\mathcal{I}_E^*} \exp\Big(n(i(E))\sum_{F\subseteq_\ominus E}\theta(i_F)\Big)\ p_\emptyset^{n(i^*)+\sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}_E^*} n(i(E))}, \\
&= p_\emptyset^N \exp\sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}_E^*}\Big(n(i(E))\sum_{F\subseteq_\ominus E}\theta(i_F)\Big), \\
&= p_\emptyset^N \exp\sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}_E^*} n(i_E)\theta(i_E), \\
&= \exp\{\sum_{E\in\mathcal{E}_\ominus}\sum_{i_E\in\mathcal{I}_E^*}\theta(i_E)n(i_E) + N\theta_\emptyset\},\ .
\end{aligned}$$

where the second equality is due to (3), the third to the fact that $\exp\theta_\emptyset = p_\emptyset$, the fourth to the identification of the exponent of $p_\emptyset$ as the total count N, the fifth to the definition of marginal counts $n(i_E)$ and the sixth to $p_\emptyset = \exp\theta_\emptyset$. Finally (8) follows from (4), (6) and Lemma 3.1.