J. Statist. Comput. Simul., 2000, Vol. 68, pp. 23–37
Reprints available directly from the publisher
Photocopying permitted by license only

STOCHASTIC SEARCH VARIABLE SELECTION FOR LOG-LINEAR MODELS

IOANNIS NTZOUFRAS[a,*], JONATHAN J. FORSTER[b,†] and PETROS DELLAPORTAS[c,‡]

[a]*Department of Statistics, Athens University of Economics and Business, Greece;* [b]*Department of Mathematics, University of Southampton, UK;* [c]*Department of Statistics, Athens University of Economics and Business, Greece*

(Received 23 September 1998; In final form 10 April 2000)

We develop a Markov chain Monte Carlo algorithm, based on 'stochastic search variable selection' (George and McCulloch, 1993), for identifying promising log-linear models. The method may be used in the analysis of multi-way contingency tables where the set of plausible models is very large.

Keywords: Bayesian analysis; Contingency table; Gibbs sampling; Markov chain Monte Carlo

1. INTRODUCTION

For contingency table data, log-linear models provide a convenient way of investigating relationships between the categorical variables by which individuals have been cross-classified. Suppose that the data are in the form of a t-way $m_1 \times m_2 \times \cdots \times m_t$ contingency table. The total number of cells in the table is $n = \prod_i m_i$ and the cell counts $y_1, \ldots, y_n$ are assumed to be realisations of independent Poisson random

*e-mail: jbn@stat-athens.aueb.gr
†Corresponding author. e-mail: jjf@maths.soton.ac.uk
‡e-mail: petros@aueb.gr

variables $Y_1, \ldots, Y_n$ with corresponding cell means $\mu_1, \ldots, \mu_n$. A saturated Poisson log-linear model assumes that

$$\log \mu = X\boldsymbol{\beta}$$

where $\mathbf{X}$ is an $n \times n$ matrix of full rank. Suppose that

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$$

where $\mathbf{X}_j$ is a $n \times n_j$ matrix with corresponding parameter (subvector) $\boldsymbol{\beta}_j$. Then

$$\log \boldsymbol{\mu} = \sum_j \mathbf{X}_j \boldsymbol{\beta}_j$$

and non-saturated log-linear models may be specified by setting certain $\boldsymbol{\beta}_j$ equal to $\mathbf{0}$.

The identification of plausible non-saturated log-linear models is equivalent to a 'variable selection' problem, and we can introduce binary indicator parameters $\gamma_1, \ldots, \gamma_p$ which take the value 1 if a term is present in the model ($\boldsymbol{\beta}_j$ is non-zero) and 0 otherwise. A Bayesian approach to variable selection requires the marginal posterior distribution of $\boldsymbol{\gamma}$

$$f(\boldsymbol{\gamma}|\mathbf{y}) \propto f(\boldsymbol{\gamma}) \int f(\mathbf{y}|\boldsymbol{\gamma}, \boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{\gamma}) d\boldsymbol{\beta}, \tag{1}$$

assuming that the prior distributions $f(\boldsymbol{\beta}|\boldsymbol{\gamma})$ are all proper. For log-linear models, the integral on the right hand side cannot usually be calculated directly, and analytic or Monte Carlo approximations are required.

2. STOCHASTIC SEARCH VARIABLE SELECTION

Stochastic Search Variable Selection (SSVS) was originally proposed by George and McCulloch (1993) as a Markov chain Monte Carlo method for variable selection in the linear regression model $\boldsymbol{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$. The SSVS approach assumes that the $n \times p$ matrix $\mathbf{X}$ contains the values of *all* potential explanatory variables and $\boldsymbol{\beta}$ is a vector of fixed dimension p for all models. SSVS differs from the

standard modelling approach as no term is ever completely absent from the model, and 'absence' of a term here implies that the corresponding parameter is constrained to be 'close to zero' so that the effect of the term is insignificant. Note that this constraint is a feature that characterizes the SSVS approach. The advantage of using a Gibbs sampler on the full parameter space of dimension p is compensated by the fact that careful choice of c_j^2 is required. An alternative approach is to use a point mass at 0, instead of a $N(0, \tau_j^2)$ density (Mitchell and Beauchamp, 1988). However, then the Gibbs sampler cannot be used and alternative MCMC approaches may be required (see Dellaportas *et al.*, 1998).

For linear regression models, George and McCulloch (1993) construct the prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in two stages. Firstly, the regression parameters β_j; $j = 1, \ldots, n$ are given normal prior distributions conditional on the model indicators γ_j

$$\beta_j | \gamma_j \sim (1 - \gamma_j) N(0, \tau_j^2) + \gamma_j N(0, c_j^2 \tau_j^2), \quad j = 1, \ldots, n. \tag{2}$$

The prior parameters τ_j^2 and c_j^2 are chosen so that τ_j^2 is 'small' and $c_j^2 \tau_j^2$ is 'large'. Precise values for these parameters are discussed in Section 3. Hence, if $\gamma_j = 1$ and the term is considered to be present in the model, then the prior distribution for the corresponding parameter β_j is vague, and its posterior distribution will largely be determined by the data. If $\gamma_j = 0$ and the term is considered to be absent from the model, then the prior distribution for the corresponding parameter β_j forces this parameter to be close to zero, as required. The precise value of c_j^2 depends on which values of β_j are considered to be 'effectively zero', and which are considered to be practically significant.

The second stage of the prior distribution specifies $f(\gamma)$ the marginal prior distribution of the model indicator variables, and is chosen to reflect prior belief in the presence or absence of particular model terms.

Posterior model probabilities based on the posterior distribution of γ calculated by SSVS will depend on $\mathbf{y}$, $f(\gamma)$, τ_j^2 and c_j^2. In particular, the prior parameter τ_j^2 is peculiar to the SSVS approach, and has no interpretation within the standard framework, where terms may be truly absent from a model, with the corresponding parameter set to zero. The balance between parsimonious and complex models is

controlled there by the equivalent of $c_j^2\tau_j^2$, the prior variances of the model parameters when the corresponding terms are included in the model. For SSVS the balance between parsimonious and complex models is controlled primarily by c_j^2, the ratio of the prior variances.

An advantage of the SSVS approach is that the parameter space is $\mathbb{R}^p$ for all models, and therefore Gibbs sampling provides a straightforward MCMC method for generating from the posterior distribution $f(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y})$. For linear regression models, George and McCulloch (1993) show that the univariate posterior conditional distributions, required for Gibbs Sampling, are easy to generate from. The conditional distributions for the β_j parameters are univariate normal, and those for the γ_j indicators are Bernoulli (see George and McCulloch, 1993 for exact details, as there is also the unknown variance σ^2 to be considered).

3. SSVS FOR LOG-LINEAR MODELS

George, McCulloch and Tsay (1995) discuss the extension of the SSVS approach to generalised linear models. Here we consider the special case where the models under consideration are hierarchical log-linear models, which have the property that if a particular interaction term is present in a model. then all marginal terms must also be present.

This class of models has three features which make it worth separate consideration. Firstly, for contingency tables involving factors with more than two levels, model terms for main effects and interactions are represented by more than one model parameter. Then, each $\boldsymbol{\beta}_j$ may be a vector and the normal distributions in (2) are multivariate. However, the principle remains the same, with two parameters controlling the prior dispersion when the term is present in, and absent from, the model.

The second special feature of the hierarchical log-linear models is that the γ_i may not be considered to be independent *a priori* as values of $\boldsymbol{\gamma}$ which correspond to non-hierarchical models are prohibited. Finally, even for contingency tables with moderate numbers of cells, the number of hierarchical log-linear models may be so great that the only feasible approach for estimating posterior model probabilities is to use a Monte Carlo approach such as SSVS.

Dellaportas and Forster (1999) discuss vague prior distributions for log-linear model parameters for multiway contingency tables. They consider the components of $\boldsymbol{\beta}_j$ to be a linearly independent subset of the usual log-linear model parameters satisfying sum-to-zero constraints. The index j now corresponds to a particular main effect or interaction, so it can be thought of as a *set* of factors. Dellaportas and Forster (1999), following Knuiman and Speed (1988), recommend that *a priori* $\text{Var}(\boldsymbol{\beta}_j) \propto \Sigma_j$ where

$$\Sigma_j = \frac{1}{n}\prod_{i \in j} n_i \bigotimes_{i \in j} \left(\mathbf{I}_{(n_i-1)} - \frac{1}{n_i}\mathbf{J}_{(n_i-1)} \right)$$

where $\mathbf{I}$ and $\mathbf{J}$ are respectively the identity matrix, and a square matrix of 1s, of appropriate dimension. This is an isotropic covariance structure which maintains the constraints. In order to adapt SSVS to problems where model terms involve multiple parameters we modify the mixture prior distribution (2) so that it involves appropriate multivariate normal components. Therefore. the prior for $\boldsymbol{\beta}_j$ takes the form

$$\boldsymbol{\beta}_j | \gamma_j \sim (1 - \gamma_j) N_{d_j}(\mathbf{0}, \tau_j^2 \Sigma_j) + \gamma_j N_{d_j}(\mathbf{0}, c_j^2 \tau_j^2 \Sigma_j) \tag{3}$$

where d_j denotes the dimension.

It is necessary to specify c_j and τ_j. For log-linear models, this task may be more straightforward than for linear regression models, as there is no intrinsic unknown scale parameter σ^2 and hence parameter values have the same interpretation in different examples. For example, for a vague prior distribution, Dellaportas and Forster (1999) chose $\text{Var}(\boldsymbol{\beta}_j) = 2n\Sigma_j$, which suggests that $c_j^2 \tau_j^2 = 2n$.

When $\boldsymbol{\beta}_j$ is one-dimensional the specification of the prior may be completed by using the method suggested by George and McCulloch (1998), of specifying the value of $|\boldsymbol{\beta}_j|$ at which the densities of the two components of the prior distribution are equal. At this value $f(\boldsymbol{\beta}_j | \gamma_j = 0) = f(\boldsymbol{\beta}_j | \gamma_j = 1)$, and therefore it may be considered as the smallest value for which the term is considered of practical significance. Consider the simplest possible example, the 2×2 contingency table. When j corresponds to the interaction between the two binary factors, then $d_j = 1$, $\Sigma_j = (1/n)$ and β_j is equal to one quarter of the log cross-product ratio. Now, suppose that δ_j is the

smallest value of β_j of practical significance. Then

$$\frac{1}{c_j\tau_j\sqrt{2\pi/n}}\exp\left(-\frac{n\delta_j^2}{2c_j^2\tau_j^2}\right)=\frac{1}{\tau_j\sqrt{2\pi/n}}\exp\left(-\frac{n\delta_j^2}{2\tau_j^2}\right)$$

which implies that

$$\delta_j=\tau_j\sqrt{\frac{2c_j^2\log c_j}{n(c_j^2-1)}}\approx\tau_j\sqrt{\frac{2\log c_j}{n}}.$$

Therefore the prior may be constructed by specifying δ_j and either τ_j or c_j. If, as suggested above, $c_j^2\tau_j^2=2n$, then the prior is specified through

$$\delta_j=2\sqrt{\frac{\log c_j}{c_j^2-1}}\approx\frac{2\sqrt{\log c_j}}{c_j}.$$

Where the model consists completely of one-dimensional parameters, this approach can be adopted for every model term j. Indeed, for a reference analysis, it may be appropriate to choose the same values for c_j and τ_j for every term j. For example, if $c_j=10^3$ and $\tau_j=2\sqrt{2}\times10^{-3}$, then $c_j^2\tau_j^2=8$ as suggested by Dellaportas and Forster, and $e^{4\delta_j}\approx1.021$, so the boundary between significant and insignificant cross product ratios would be represented by an increase in 'risk' of around 2.1%. The relationship between c and δ for the 2×2 reference example is illustrated in Figure 1.

For multidimensional $\boldsymbol{\beta}_j$ parameters, some adjustment may be required. For example, it is possible to ensure that the ratio of the two components of the mixture prior density at $\boldsymbol{\beta}_j=\mathbf{0}$ is invariant to the dimension d_j of $\boldsymbol{\beta}_j$ by setting $\log c_j$ proportional to $1/d_j$. Then $f(\gamma_j=1|\boldsymbol{\beta}_j=\mathbf{0})$ will not depend on the dimension d_j of $\boldsymbol{\beta}_j$. This seems to ensure sensible results in problems where the d_j vary, and is intuitively plausible as the interpretation of $\boldsymbol{\beta}_j=\mathbf{0}$ is invariant to the dimension d_j.

An alternative approach for multidimensional $\boldsymbol{\beta}_j$ is to adapt the approach of George and McCulloch described above, and to choose c_j and τ_j by again considering the values of $\boldsymbol{\beta}_j$

$$\boldsymbol{\beta}_j^T\Sigma_j^{-1}\boldsymbol{\beta}_j=c_j^2\tau_j^2\frac{2d_j\log c_j}{c_j^2-1}$$

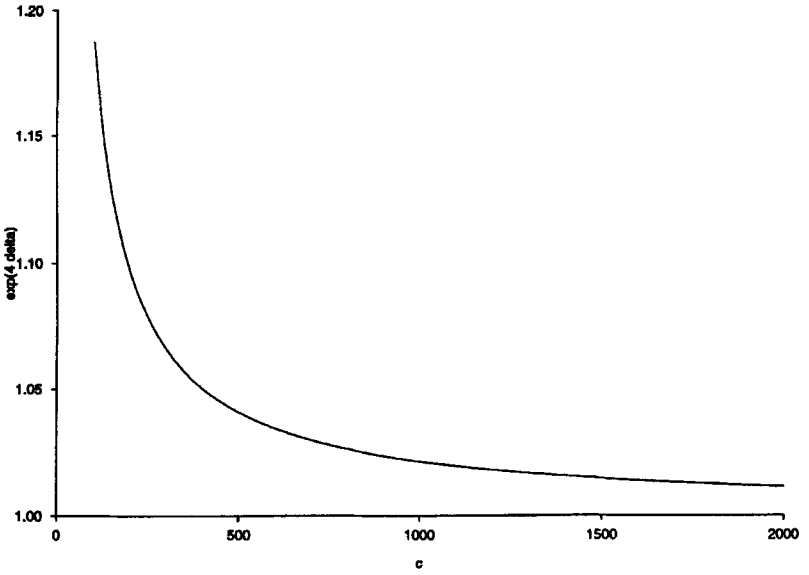

FIGURE 1 The relationship between cross-product ratio boundary ($=e^{4\delta}$) and c for the 2×2 table.

where the two components of the mixture prior densities have equal values. Suppose that

$$I_j = \left\{ \boldsymbol{\beta}_j : \boldsymbol{\beta}_j^T \Sigma_j^{-1} \boldsymbol{\beta}_j \leq c_j^2 \tau_j^2 \frac{2d_j \log c_j}{c_j^2 - 1} \right\}$$

denotes the 'region of insignificance'. Then it is possible to determine c_j and τ_j so that $P(\boldsymbol{\beta}_j \in I_j | \gamma_j = 0)$ is the same for all j, regardless of the value of d_j. We have

$$P(\boldsymbol{\beta}_j \in I_j | \gamma_j = 0) = F_{d_j}\left(c_j^2 \frac{2d_j \log c_j}{c_j^2 - 1} \right)$$

where F_d is the distribution function of a chi-squared random variable with d degrees of freedom. Therefore, if c is the value of c_j proposed when $d_j = 1$, then setting these probabilities equal at dimensions 1 and d_j we obtain the corresponding value for $d_j > 1$ as the solution to the

equation

$$F_{d_j}\left(c_j^2 \frac{2d_j \log c_j}{c_j^2 - 1}\right) = F_1\left(c^2 \frac{2\log c}{c^2 - 1}\right) \tag{4}$$

and approximately as c_j^2 is large,

$$c_j = \exp\left(\frac{1}{2d_j} F_{d_j}^{-1}[F_1(2\log c)]\right). \tag{5}$$

To see how c_j varies with dimension, see Figure 2, which is a plot of $\log c_j$ against d_j when $c = 1000$. The solid line represents $\log c_j \propto 1/d_j$ while the dotted line represents the values given by (5).

All that remains is to specify the prior distribution $f(\gamma)$ of the model indicator variables which reflects prior belief in the presence or absence of particular model terms. As mentioned above. we enforce the usual restriction that all models under consideration should be hierarchical. We assume that all combinations of γ_j terms which

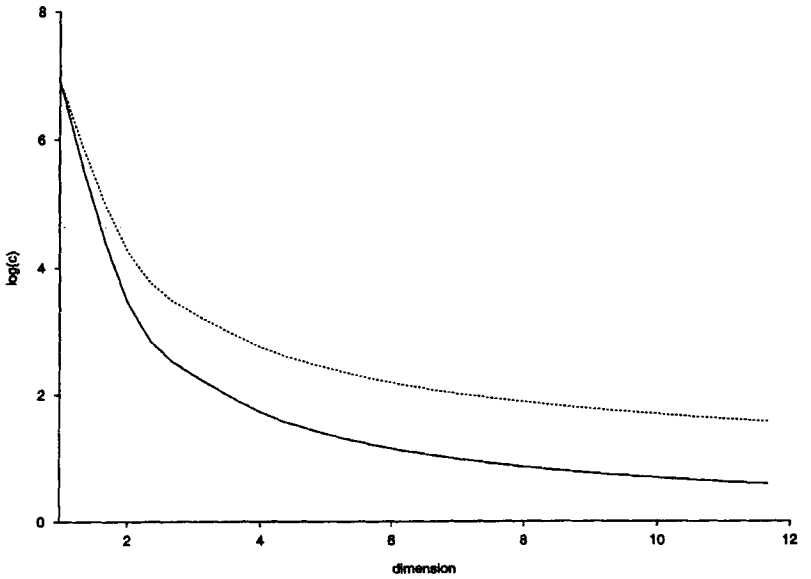

FIGURE 2 The relationship between $\log c_j$ and d_j for $c = 1000$.

correspond to hierarchical models are *a priori* equally likely. This prior distribution could easily be modified if required.

For log-linear models, and the prior distributions specified above, the univariate conditional distributions for the γ_j indicators take exactly the form given by George and McCulloch (1993). They are Bernoulli with

$$P(\gamma_j = 1|\boldsymbol{\beta}, \gamma_{\setminus j}) = \frac{f(\beta_j|\gamma_j = 1)f(\gamma_j = 1|\gamma_{\setminus j})}{f(\beta_j|\gamma_j = 1)f(\gamma_j = 1|\gamma_{\setminus j}) + f(\beta_j|\gamma_j = 0)f(\gamma_j = 0|\gamma_{\setminus j})} \quad (6)$$

assuming that models including and excluding the corresponding term are still hierarchical with all other γ_j at their current values; otherwise γ_j remains at its current value. This Bernoulli probability involves only the calculation of the normal prior probabilities of the current value of β_j for both possible values of γ_j.

The univariate conditional distributions of the parameters β_j, or components of $\boldsymbol{\beta}_j$ if it is a vector, take the form

$$f(\beta_j|\boldsymbol{\beta}_{\setminus j}, \gamma, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\alpha_j^2\beta_j^2 + \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \sum_{i=1}^{n} \exp[\mathbf{X}\boldsymbol{\beta}]_i \right), \quad j = 1, \ldots, n \quad (7)$$

where α_j is a constant depending on γ_j and on the (multivariate) normal prior for β_j. Although this univariate conditional density is not of any recognised form, it is log-concave, and therefore can be generated from efficiently using the adaptive rejection sampling algorithm proposed by Gilks and Wild (1992).

Therefore, in principle, we can use the Gibbs sampler to generate from the full posterior distribution $f(\gamma, \boldsymbol{\beta}|\mathbf{y})$ and the posterior model probabilities $f(\gamma|\mathbf{y})$ can be estimated by the sample proportions for γ, with the most promising models corresponding to the most frequently observed γ. The generating procedure of the SSVS algorithm can be summarised as follows:

- Generate β_j, for $j = 1, \ldots, n$, from the conditional posterior density given by (7) using the adaptive rejection sampling algorithm proposed by Gilks and Wild (1992).
- Generate γ_j, for $j = 1, \ldots, n$, from a Bernoulli distribution with success probability given by (6).

4. EXAMPLES

We present two examples, one where alternative methods of calculating $f(\gamma|\mathbf{y})$ are available, and another where MCMC methods are required.

4.1. Example One: $3 \times 2 \times 4$ Contingency Table

This example is a $3 \times 2 \times 4$ contingency table presented by Knuiman and Speed (1988) where 491 individuals are classified by three categorical variables: obesity (O: low, average, high), hypertension (H: yes, no) and alcohol consumption (A: 0, 1−2, 3−5, 6 + drinks per day).

The results are summarised in Table I. There are nine possible hierarchical models in total, but the data strongly favour the model of mutual independence of H, O and A, with some evidence of an interaction between O and H. For comparison with our MCMC results, we also present the model probabilities obtained using Laplace's method, an analytic approximation.

SSVS and Laplace's method provide similar approximations to the posterior model probabilities, although the latter method seems to underestimate the probability of the more complex model OH + A. The presence of these two models, with the mutual independence model dominant is also a feature when these data are analysed using the 'glib' procedure provided by Raftery (1996).

4.2. Example Two: A 2^6 Contingency Table

Consider the 2^6 table of risk factors for coronary heart disease presented by Edwards and Havránek (1985). This table has also been analysed by Madigan and Raftery (1994) and Madigan *et al.* (1995)

TABLE I Posterior model probabilities estimated by SSVS and Laplace's method ($c = 1000$)

	log $c_j \propto 1/\delta_j$		*Equation* (5)	
Model	*SSVS*	*Laplace*	*SSVS*	*Laplace*
O + H + A	0.9541	0.9825	0.9167	0.9774
OH + A	0.0431	0.0145	0.0827	0.0220
Others	0.0028	0.0030	0.0006	0.0006

using both stepwise and MCMC Bayesian model selection algorithms for decomposable log-linear models. Decomposable models are a subset of the hierarchical models which we consider in this paper. However, there are many interesting models which are not decomposable, and therefore we present the results of our SSVS hierarchical model selection approach. Here, the six variables are: A, smoking; B, strenuous mental work; C, strenuous physical work; D, systolic blood pressure; E, ratio of α and β lipoproteins; F, family anamnesis of coronary heart disease. The Gibbs sampler was extremely mobile, and gave good estimates of the model probabilities in a reasonably short time. It is interesting to note that the two models with highest probability are the same as those determined by Edwards and Havránek (1985) using their procedure for identifying acceptable parsimonious models. However, their procedure could not specify the uncertainty associated with this model selection, something which SSVS is clearly capable of. Furthermore, none of these models are decomposable, and hence they were not identified by Madigan and Raftery (1994) or Madigan *et al.* (1995).

The posterior model probabilities displayed in Table II indicate that, no single model is dominant. The associated model probabilities appropriately summarise model uncertainty in light of the observed data. Quantification of model uncertainty through posterior model probabilities is a natural advantage of a Bayesian approach because inferences and predictions may be appropriately adjusted. For example, the predictive density of any parameter of interest. say z, is given by

$$f(z|\mathbf{y}) = \sum_{\gamma} f(z|\gamma, \mathbf{y}) f(\gamma|\mathbf{y})$$

where $f(\gamma|\mathbf{y})$ is given by (1). For further discussion, see Draper (1995).

TABLE II SSVS posterior model probabilities ($c = 1000$)

Model	*Probability*[1]	*Probability*[2]
AC + BC + AD + AE + CE + DE + F	0.2642	0.2686
AC + BC + AD + AE + BE + DE + F	0.1604	0.1599
AC + BC + AD + AE + BE + CE + DE + F	0.0691	0.0713
AC + BC + AD + AE + CE + DE + BF	0.0655	0.0701

[1] Results from sample of one million iterations and 30,000 burnin.
[2] Results from 25 independent samples of 42,000 iterations.

We need to be sure that the model probabilities are accurate, and that there are no highly probable models to which the transition probability within a reasonable number of iterations of the SSVS algorithm is very small. In particular, those models which have not been visited by the Gibbs sampler, and whose probabilities are therefore estimated to be zero, should indeed have negligible probability. This is particularly relevant in examples such as the present one where the number of possible models (almost 8 million) exceeds the length of the Gibbs sampler chain.

For log-linear models, Dellaportas and Forster (1999) proposed restarting an MCMC algorithm from a number randomly chosen points in model space, and observing whether the most probable models are visited within a small number of iterations. Observing consistency across parallel independent runs of a Gibbs sampler is a convenient way of reassuring ourselves that the model probabilities are accurate. In the current example, we used the same strategy as in Dellaportas and Forster (1999). Our estimate of the posterior distribution of γ is based on 1,000,000 iterations (after discarding the first 30,000 iterations as burn-in); see Table II. Next, we split the sample into 25 equal batches and obtained estimates of the posterior model probabilities based on 40,000 sample points. The results are illustrated in Figure 3 and indicate that Monte Carlo error is within satisfactory levels. Finally, following Dellaportas and Forster (1999), we generated 25 random models as starting points and we ran 25 independent parallel chains for 42,000 iterations producing the results illustrated in Figure 4 and reported in Table II. Again. there is evidence that all samplers have visited the region of higher posterior mass. Finally, a final check to reassure ourselves that SSVS is quick in locating the higher posterior regions is to examine how fast our algorithm visits the four most probable models. 90% of the samplers visited the most probable models in less than 3,000 iterations and 70% in less than 1,900 iterations.

George and McCulloch (1997) investigated the equivalent problem in normal linear regression settings and they suggested either an exhaustive search of the model space (if possible), or construction of faster MCMC algorithms which exploit the ability to analytically integrate out all parameters but γ_j.

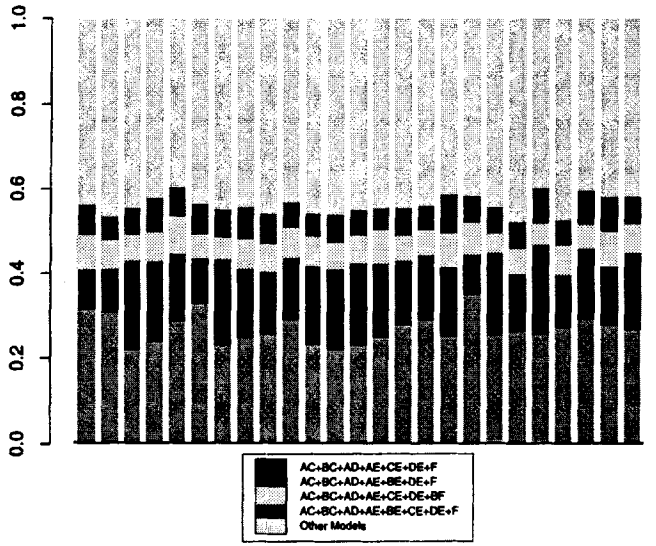

FIGURE 3 Posterior model probabilities estimated with 1,000,000 iterations divided in 25 equal batches.

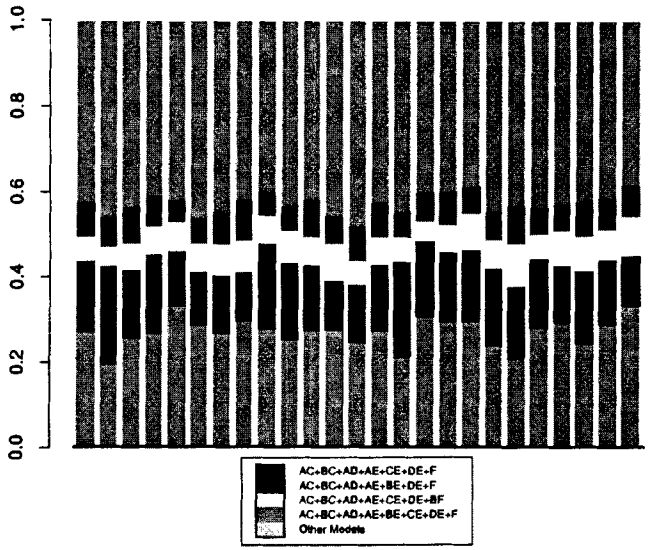

FIGURE 4 Posterior model probabilities estimated with 25 batches of 42,000 iterations starting from 25 randomly chosen models.

5. DISCUSSION

Stochastic search variable selection is extremely efficient for identifying promising log-linear models. For a six-way contingency table there are approximately 7.8 million possible hierarchical log-linear models, so direct evaluation or approximation of model probabilities using (1) would be extremely time consuming. For tables with seven or more factors, such a calculation would be completely infeasible.

In many practical situations, where non-saturated models are considered implausible, the philosophy of the SSVS approach, focussing on practical significance may be extremely attractive. Certainly, as the number of observations in the table increases, then the standard approach will tend to favour more complex models. By choosing suitable values for the prior parameters c_j^2 for SSVS, reasonably parsimonious models may still be selected. Obviously, where possible, these prior parameters should be chosen by carefully considering each of the log-linear model parameters $\boldsymbol{\beta}_j$. However, where there are very many of these, or a reference analysis is required, the methods for choosing c_j^2 suggested in Section 3 seem to provide sensible results. Similarly, we have chosen to give all models equal prior probability. However, where relevant prior information exists, it may be possible to rule out certain models as implausible, hence reducing the size of the set of possible models.

Acknowledgements

We are grateful to an associate editor and a referee for useful comments.

References

Dellaportas, P. and Forster, J. J. (1999) Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-linear Models. *Biometrika*, **86**, 615–633.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (1998) On Bayesian Model and Variable Selection Using MCMC. *Technical Report*, Department of Statistics, Athens University of Economics and Business.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. Roy. Statist. Soc. B*, **57**, 45–70.

Edwards, D. and Havránek, T. (1985) A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika*, **72**, 339–351.

George, E. I. and McCulloch, R. E. (1993) Variable Selection *via* Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339–373.

George, E. I., McCulloch, R. E. and Tsay, R. S. (1995) Two Approaches to Bayesian Model Selection with Applications. In: Berry, D. A., Chaloner, K. M. and Geweke, J. K. (Eds.). *Bayesian Analysis in Statistics and Econometrics*. Wiley, New York, pp. 339–348.

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

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

Kuo, L. and Mallic, B. (1988) Variable selection for regression models. *Sankhya B*, **60**, 65–81.

Madigan, D. and Raftery, A. E. (1994) Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Madigan, D., Raftery, A. E., York, J., Bradshaw, J. M. and Almond, R. G. (1995) Strategies for Graphical Model Selection. In: Cheesman, P. and Oldford, R. W. (Eds.). *Selecting Models from Data: AI and Statistics IV*. Springer-Verlag, New York, pp. 91–100.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, **83**, 1023–1036.

Raftery, A. E. (1993) Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, **83**, 251–266.