Biometrika (2001), **88,** 2, *pp.* 317–336

Printed in Great Britain

On the Bayesian analysis of population size

BY R. KING

School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, U.K.

ruth.king@bris.ac.uk

AND S. P. BROOKS

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, U.K.

s.p.brooks@statslab.cam.ac.uk

SUMMARY

We consider the problem of estimating the total size of a population from a series of incomplete census data. We observe that inference is typically highly sensitive to the choice of model and we demonstrate how Bayesian model averaging techniques easily overcome this problem. We combine and extend the work of Madigan & York (1997) and Dellaportas & Forster (1999) using reversible jump Markov chain Monte Carlo simulation to calculate posterior model probabilities which can then be used to estimate model-averaged statistics of interest. We provide a detailed description of the simulation procedures involved and consider a wide variety of modelling issues, such as the range of models considered, their parameterisation, both prior choice and sensitivity, and computational efficiency. We consider a detailed example concerning adolescent injuries in Pennsylvania on the basis of medical, school and survey data. In the context of this example, we discuss the relationship between posterior model probabilities and the associated information criteria values for model selection. We also discuss cost-efficiency issues with particular reference to inclusion and exclusion of sources on the grounds of cost. We consider a decision-theoretic approach, which balances the cost and accuracy of different combinations of data sources to guide future decisions on data collection.

Some key words: Census data; Contingency table; Cost-effectiveness; Decision theory; Log-linear model; Markov chain Monte Carlo; Posterior model probability; Reversible jump.

1. INTRODUCTION

Many areas of scientific research are concerned with the estimation of population size. Data are often available only in the form of incomplete population counts and it is through these that an estimate of the total population size is required. However, the total number of individuals observed from all sources will underestimate the true population size.

Under the assumption that individuals recorded by any source are uniquely identifiable, a contingency table can be constructed from the data. The cells of the table refer to the numbers of individuals that appear in each dataset combination of the sources. For example the cell $(1, 1, 0)$ might correspond to the number of individuals observed by the first two of three sources, but not the last, with 1/0 denoting presence/absence on a list.

An unbiased estimator for the true population size is obtained via the estimation of the missing cell (0, 0, 0) corresponding to individuals who belong to the population, but are not observed by any of the sources.

There have been many classical analyses of such data, for example Fienberg (1972), Edwards & Havránek (1985), Hook et al. (1980) and Hook & Regal (1995). These analyses typically use log-linear models, finding the model which provides the 'best' fit to the data, using likelihood ratio tests and/or information criteria. However, Rasch models and more general latent class models may also be used; see Fienberg et al. (1999) and § 7 for further discussion of these models. Once we have selected the model, the total population is then estimated using the maximum likelihood estimator for the missing cell, combined with the observed number of individuals. There are two problems with this approach. First, as the number of sources grows, the number of possible models grows exponentially and it quickly becomes impossible to discriminate between all models within such a wide class. Secondly, the maximum likelihood estimator for the missing cell, and hence the estimate of the total population size, is often highly sensitive to the choice of model. Thus, it is often difficult to ascribe much confidence to the results obtained.

The Bayesian approach overcomes these problems by using simulation methods to calculate the posterior probability of each model. These posterior model probabilities may then be used to obtain a model-averaged estimate of the population size, overcoming the model-dependence problem of the classical approach. Madigan & York (1997) propose a method known as Markov chain Monte Carlo model composition to obtain these posterior model probabilities. A substantial drawback of their simulation method is that it does not simultaneously explore both parameter and model space and thus requires a separate algorithm to obtain estimates of both the parameters and the missing cell. Additionally, because of the way in which the prior for the cell probabilities is structured, only decomposable models can be considered.

Dellaportas & Forster (1999) propose an alternative method for obtaining posterior model probabilities for the analysis of complete contingency tables using log-linear models. They use multivariate normal priors for the log-linear parameters and reversible jump Markov chain Monte Carlo simulation techniques for exploring both model and parameter space simultaneously. Such methods can be used to consider all hierarchical models and so overcome the restriction to decomposable models of the analysis of Madigan & York (1997).

The aim of the present paper is to combine and extend these two approaches. We shall discuss how reversible jump Markov chain Monte Carlo simulation methods may be applied when there is a missing cell within the table. We consider the class of hierarchical log-linear models and discuss various parameterisation issues and their effect on computational efficiency. In § 2 we establish the notation to be used throughout the paper before discussing the issues involved in adopting a Bayesian approach to the population undercount problem. In § 3 we discuss the issue of prior specification and its implications for computational efficiency. Implementation issues are described in § 4 and, in particular, methods for improving Markov chain Monte Carlo mixing are discussed. Section 5 provides an example illustrating the power and flexibility of our approach. We compare our approach to that of a classical analysis and investigate the relationship between posterior model probabilities and standard information criteria for model discrimination. In § 6, we discuss the question of cost efficiency of data sources by comparing the cost involved in collecting the data with the additional accuracy they provide. We conclude with further discussion of the context of this work and suggest areas for future research.

2. Notation for log-linear models

2·1. *Source and cell identification*

We assume that the data are provided in the form of individuals uniquely identified as having been observed by one or more of a set of sources S, where $|S|$ denotes the number of sources. Each source has two levels, 0 and 1, representing the absence/presence of individuals from the corresponding list. We denote the set of levels for each source $\gamma \in S$ by K_γ which, in our case, is simply $\{0, 1\}$ for all $\gamma \in S$. In order to index the cells within the contingency table, we define $K = \prod_{\gamma \in S} K_\gamma = \{0, 1\}^{|S|}$; the cells of the contingency table can then be indexed by $k \in K$, with corresponding cell probabilities and counts given by p_k and n_k respectively. We also define $\mathscr{P}(S)$ to be the power set, or set of subsets, of S, that is $\mathscr{P}(S) = \{s : s \subseteq S\}$. Note that we shall allow for a constant log-linear term by including the empty set in $\mathscr{P}(S)$, that is $\varnothing \in \mathscr{P}(S)$. For example, consider three sources, A, B and C. Then

$$S = \{A, B, C\}, \quad \mathscr{P}(S) = \{\varnothing, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\},$$

$$K_\gamma = \{0, 1\} \quad (\gamma = A, B, C),$$

$$K = \{0, 1\}^3 = \{(k_1, k_2, k_3) : k_i \in \{0, 1\} \text{ for all } i = 1, 2, 3\}$$

$$= \{(0, 0, 0), (0, 0, 1), (0, 1, 0), \ldots, (1, 1, 1)\}.$$

Models can be indexed by m for any $m \subseteq \mathscr{P}(S)$, where the set m denotes the terms to be included in the model; for example $m = \{\varnothing, \{A\}, \{B\}, \{C\}, \{A, B\}\}$ denotes the presence of a constant term, the three main effects and an AB interaction term.

2·2. *Log-linear parameters*

Given model $m \subseteq \mathscr{P}(S)$, we denote the associated log-linear parameter vector by $\theta^m = (\theta^c : c \in m)$. The cell probabilities can then be expressed, up to a normalisation constant, as

$$\log p_k = \sum_{c \in m} I^c(k)\theta^c \quad (k \in K),$$

where the $I^c(k) = \pm 1$ are sign functions ensuring that the usual conditions for identifiability are observed. An explicit expression for these sign functions is given in King & Brooks (2001).

To clarify this notation we take the three-source example above and consider the cell $k = (0, 0, 1)$ corresponding to individuals that were only observed by source C. We consider model m with log-linear parameters representing sources A, B and C and the two-way interaction between sources A and C, that is $m = \{\varnothing, \{A\}, \{B\}, \{C\}, \{A, C\}\}$. Then $\theta^m = \{\theta^\varnothing, \theta^A, \theta^B, \theta^C, \theta^{AC}\}$, with corresponding sign function values $I^\varnothing = I^A(k) = 1 = I^B(k)$ and $I^C(k) = -1 = I^{AC}(k)$, with $k = (0, 0, 1)$. Therefore, the corresponding cell probability is given by

$$\log p_{001} = \theta\varnothing + \theta^A + \theta^B - \theta^C - \theta^{AC}.$$

For the remainder of this paper, we shall restrict our attention to the class of hierarchical models, that is models $m \subseteq \mathscr{P}(S)$ such that $c \in m \Rightarrow d \in m$, for all $d \subseteq c$. For example, if $c = \{A, B\} \in m$, then $\{A\}, \{B\}, \varnothing \in m$. We further restrict our attention to models that always contain the constant term $\varnothing$ and each of the main effects, so that the vector θ^m always contains the parameters $\{\theta^c : c \in S\}$. Finally, we exclude the saturated model. We

use the standard shorthand notation to specify hierarchical log-linear models, by listing only the maximal elements within the model, with all other terms necessarily implied. In addition, we abbreviate the elements of a model in the usual way so that we express $c = \{A, B\}$ as $c = AB$, for example.

To complete the model specification, we assume that given a total population size, N, the vector of the number of observations in each of the table cells, n, has a multinomial distribution with parameters N and p, the latter being the vector of cell probabilities.

3. PRIOR SPECIFICATION

The parameter of primary interest is that representing the size of the total population, N. Madigan & York (1997) consider the prior $N|\lambda \sim \text{Po}(\lambda)$, where λ is unknown and has a hyperprior with fixed parameters. A $\Gamma(\alpha, \beta)$ is chosen to be the corresponding prior for λ, where α and β are to be specified. This Poisson/gamma prior is overdispersed relative to the Poisson, and is a more general form of the negative binomial distribution, but it is more interpretable because of the hierarchical structure. In the case where no prior information exists, the gamma hyperparameters should reflect this by possessing a large degree of variation. Alternatively the Jeffreys prior may be used in which the prior distribution for N is proportional to $1/N$. Fienberg et al. (1999) adopt a family of priors proportional to N^{-c} for some positive constant c. Priors of this form tend to have somewhat heavier tails than the simple Poisson/gamma prior above and may be appropriate in some situations, depending upon the prior beliefs that are to be expressed.

A more informative prior may be appropriate in certain circumstances. For example, prior opinion may be that the sources are fairly exhaustive in identifying individuals, in which case a prior for N with a mean close to the total observed population may be adopted. Uncertainty as to the most likely proportion of the total population actually observed may be reflected by a correspondingly large standard deviation. The total cell count can also be incorporated into the log-linear parameters themselves, where we assume that the cell counts are of log-linear form rather than the cell probabilities. However, this increases the complexity of expressing a prior, when there is prior information concerning the total population.

In practice, and particularly when available data are sparse, it is important to check for sensitivity to these prior assumptions in terms of both the posterior estimates for N and also other statistics of interest such as the posterior model probabilities. In our experience, even with fairly small tables, the inference obtained is generally fairly insensitive to the choice made and we shall observe this when we come to our example in § 5.

It is also necessary to specify priors for the cell probabilities within the contingency table or, equivalently, the log-linear parameters. Madigan & York (1997) adopt the hyper-Dirichlet priors of Dawid & Lauritzen (1993) for the cell probabilities. It is particularly easy to elicit priors of this form if prior knowledge is available in the form of the effectiveness of individual sources, or combinations of sources, in capturing individuals. An advantage of this form of prior is that it allows a factorisation of the likelihood through the identification of cliques within the corresponding model graph. In a University of Pavia technical report, P. Giudici, P. J. Green and C. Tarantola illustrate how this decomposition leads to local computations for the cell probabilities, improving the computational efficiency of individual parameter updates. However, there are also several disadvantages. The first is that, by adopting a prior of this form, we must restrict ourselves to decomposable models, since priors are placed on the cliques of the model. For example, Dellaportas

& Forster (1999) provide a simple 2^6 contingency table example where the posterior model probabilities of hierarchical log-linear models, graphical models and decomposable models are compared. The corresponding results show that the most probable log-linear model is approximately 700 times more likely than the most probable decomposable model. Secondly, the hyper-Dirichlet prior requires a very large number of hyperparameters to be specified, which must be hyperconsistent and ideally compatible across all models. This is often achieved by assuming that the distributions on any clique are obtained by marginalisation from a unique distribution on a complete graph, but it certainly adds an additional level of complexity to the prior specification problem. Finally, when we use Markov chain Monte Carlo, though each parameter update involves only local computation there is additional expense both in terms of keeping track of the cliques and in updating all $2^{|S|}$ cell probabilities at each iteration.

An alternative approach is to specify a prior distribution on the log-linear parameters. It is particularly easy to elicit priors of this form if prior knowledge is in the form of the existence and/or sign of correlations between the different sources. One possible way of incorporating such information would be to place a prior with nonzero mean on the corresponding log-linear term. The corresponding variance of the term would convey the strength of the belief. An advantage of this form of prior is the conceptual and computational simplicity as well as the fact that we are no longer restricted to decomposable models. The main disadvantage is that there is no longer a decomposition of the likelihood, so that log-linear parameter updates involve global rather than local computations. Thus, there is a trade-off between a large number of local computations when using a hyper-Dirichlet prior and typically far fewer 'global' computations for priors specified on the log-linear parameters. However, the cell probabilities are updated locally within each clique, so that any comparison of computational efficiency is likely to be problem dependent. Thus, in practice, the preference for either prior should depend solely upon the form of the prior information available. If knowledge is available in the form of expected effectiveness of individual sources, then a prior on the cell probabilities would be preferred. However, if prior knowledge is expressed in terms of the existence, or otherwise, of correlations between sources, then a prior on the log-linear parameters would be more appropriate. In practice, we find that the latter case is the more common.

A typically vague prior to place on the log-linear parameters is that of a multivariate normal. With such a prior, King & Brooks (2001) calculate a practical formulation of the corresponding distribution parameters for the cell probabilities. Clearly a multivariate normal prior on the log-linear parameters induces a constrained multivariate log-normal prior on the cell probabilities. We can also express the cell probabilities as an additive logistic normal distribution (Aitchison, 1986, Ch. 6) on the $2^{|K|}-1$ simplex, with the distribution parameters given in King & Brooks (2001). However, the moments of such a distribution are analytically intractable, yet we are able to calculate easily the moments of the ratios, or log-ratios, of the cell probabilities using standard multivariate theory. It can also be shown that a vague prior on the log-linear parameters is also vague on the ratio of cell probabilities, so that the prior is consistent across these parameterisations. An uninformative prior on the log-linear parameters can be obtained by setting their means and covariances to be finite with their variances tending to infinity. Using similar results to those derived in King & Brooks (2001), we can show that the mean of the ratio of any two cell probabilities is flat over the positive real line. Further, the asymptotic correlation matrix of two cell probability ratios, with common denominator or numerator,

tends to the identity matrix as the prior variances for the log-linear parameters approach infinity. Thus, in the limit, we obtain a flat distribution over the space of cell probabilities.

This relationship between the priors for the log-linear parameters and cell probabilities may assist in the elicitation of expert priors if there is also some knowledge of the relationships between cell probabilities, as well as the interaction between sources. For example, priors may be specified for the log-linear parameters and the corresponding prior for the cell probabilities referred to the expert for approval (O'Hagan, 1998). In the absence of strong prior beliefs, both μ and Σ could themselves have hyperpriors. For illustration we shall assume, for the remainder of this paper, that $\mu = 0$ and that $\Sigma = \sigma^2 I$, where $\sigma^2 \sim \Gamma^{-1}(a, b)$ a priori.

Finally, we must specify a prior distribution for the model. Throughout this paper, we assume that all models which contain all of the main effects, excluding the saturated model, are equally likely a priori. Alternative priors which penalise models with large numbers of parameters or higher-order interactions may also be plausible in different situations. For any particular application, expert opinion should always be solicited before a decision is made.

4. IMPLEMENTATION

4·1. *Preamble*

Once the priors have been specified for all of the parameters in the model, the posterior distribution can be calculated, up to proportionality. However, this distribution is both high-dimensional and complex so we have to adopt a Markov chain Monte Carlo approach in order to explore and summarise the distribution. To update the model parameters we use a Gibbs step whenever the corresponding full conditional distribution is of a standard form, or we choose a random walk Metropolis update (Brooks, 1998). The moves between models are performed by reversible jump Markov chain Monte Carlo.

4·2. *Within model moves*

Conditioning on a particular model m, and in order to update the corresponding model parameters, we require the posterior conditional distributions. Under the priors proposed in § 3, the joint posterior can be expressed as

$$\begin{aligned}\pi_m(n_0, \lambda, \theta^m, \sigma^2 | n_{\text{obs}}) &\propto L_m(p; n) p(n_0|\lambda) p(\lambda) p(\theta^m|\sigma^2) p(\sigma^2) \\ &\propto \frac{\prod_{k \in K} p_k^{n_k} (N!)}{n_0!} \frac{\exp(-\lambda)\lambda^N}{N!} \lambda^{\alpha - 1} \exp(-\beta\lambda) \\ &\quad \times \prod_{c \in m} \left\{ \frac{1}{\sigma} \exp\left(-\frac{(\theta^c)^2}{2\sigma^2} \right) \right\} \sigma^{-2(a-1)} \exp\left(-\frac{b}{\sigma^2} \right), \qquad (1)\end{aligned}$$

where $n_{\text{obs}} = \sum_{k \in K} n_k - n_0$ denotes the sum of the observed cell counts and the vector of observations n includes the imputed value of the missing cell n_0.

From (1) it is possible to derive the form of the conditional posterior distributions of the individual parameters:

$$n_0 | \lambda, p \sim \text{Po}(\lambda p_0), \quad \lambda | n \sim \Gamma\left(\sum_{k \in K} n_k + \alpha, \beta + 1 \right),$$

$$\sigma^2 | \theta^m \sim \Gamma^{-1}\left(\frac{1}{2}|m| + a, \frac{1}{2} \sum_{c \in m} (\theta^c)^2 + b \right).$$

These can be updated individually during each iteration of the Markov chain using Gibbs updates. However, the posterior conditional distributions for the log-linear parameters are of nonstandard form:

$$\pi_m(\theta^m | n_{\text{obs}}, \sigma^2) \propto \exp\left[\sum_{c \in m}\left\{-\frac{(\theta^c)^2}{2\sigma^2} + \sum_{k \in K} n_k \theta^c I^c(k)\right\}\right]. \tag{2}$$

Thus, we use a random walk Metropolis update for these parameters. The proposal distribution for each of the parameters is Uniform, centred on the current parameter value and with a range of $2D$, so that the proposed new value of each θ^c, $c \in m$ is θ'^c, where

$$\theta'^c \sim \text{Un}(\theta^c - D, \theta^c + D).$$

The value of D is fixed and chosen via pilot tuning, though the performance of the algorithm is not greatly affected by choices of D within a wide range of reasonable values; $D \in [0{\cdot}1, 0{\cdot}3]$ appears to work well in practice for most problems. The corresponding acceptance probability for a move within the Metropolis–Hastings algorithm is

$$\alpha(\theta^c, \theta'^c) = \min\left\{1, \frac{\pi_m(\theta'^c)}{\pi_m(\theta^c)}\right\},$$

where $\pi_m(.)$ is given in (2).

4·3. *Between model moves*

Reversible jump Markov chain Monte Carlo (Green, 1995) extends the basic Markov chain Monte Carlo algorithm to deal with jumps between states of different dimensions. It is therefore ideally suited to model determination problems where we wish to explore both parameter and model space simultaneously. Suppose we propose a move of type j from a model m with parameter vector θ to a new model m' with associated parameter vector θ', where $|\theta| < |\theta'|$. The new vector θ' is generated as a function $g(\theta, u)$ of the original parameter vector and a vector of random variables from some distribution $q(u)$. Finally, suppose that the probability of proposing move type j when in model m and with parameter vector θ is given by $r(j, m, \theta)$; then the proposed move is accepted with probability.

$$\alpha(\theta, \theta') = \min(1, A),$$

where

$$A = \frac{\pi(m', \theta' | n) r(j, m', \theta')}{\pi(m, \theta | n) r(j, m, \theta) q(u)} \left|\frac{\partial \theta'}{\partial(\theta, u)}\right|, \tag{3}$$

in which $\pi(.|.)$ is the posterior distribution, given the data.

The corresponding acceptance probability of the reverse move is simply

$$\alpha(\theta', \theta) = \min\left(1, \frac{1}{A}\right); \tag{4}$$

see, for example, Richardson & Green (1997) for further discussion and implementational details.

In order to perform the reversible jump update, we consider two types of jump which we call local and global. Local jumps are those only to 'neighbouring' models, whereas

global jumps provide a mechanism for making larger jumps in the hope of improving mixing. The moves of different size are considered because of the trade-off between having a high acceptance probability for the proposed model moves and the speed with which the chain can move around the state space; local moves are generally accepted with high probabilities, but traverse the model space slowly, whereas global jumps have a lower acceptance rate, but have the potential to move the chain further from the present state.

First, we consider local jumps, proposing to jump only to hierarchical models that are neighbours of the current model. In practice, this means the addition or deletion of a single parameter, which we propose with equal probabilities. We must then determine which parameter(s) may be added/deleted. This is not entirely straightforward as we retain the restriction to hierarchical models. We begin by discussing an alternative, lexicographic indexing of the model parameters and deriving a corresponding index for the models themselves.

Let $S = \{S_1, S_2, \ldots\}$ be the ordered set of data sources. Then the log-linear parameters can be reordered by index $j(c) = 0, \ldots, 2^{|S|} - 1$, where

$$j(c) = \sum_{i:S_i \in c} 2^{i-1}.$$

We denote the jth log-linear parameter under this ordering by θ_j. The order $O(\theta_j)$, for example with each main effect of order 1, two-way interactions of order 2, etc., of the jth parameter is easily determined, since $O(\theta_j) = |i : S_i \in c|$.

For example, consider three sources, A, B and C. If we number the sources in the obvious way, so that $S_1 = A$, $S_2 = B$ and $S_3 = C$, then the log-linear parameters are lexicographically ordered as

$$(\theta_0, \theta_1, \ldots, \theta_7) \equiv (\theta^{\varnothing}, \theta^{A}, \theta^{B}, \theta^{AB}, \theta^{C}, \theta^{AC}, \theta^{BC}, \theta^{ABC}).$$

Each model is then individually identifiable with a unique value obtained by examining the parameters of which it consists so that model m is associated with the value

$$\sum_{i=0}^{2^{|S|}-1} 2^i I_m(i),$$

where $I_m(i)$ is the indicator function taking a value 1 if $c = j^{-1}(i) \in m$, and 0 otherwise.

This alternative parameterisation, or indexing, greatly improves the efficiency of the local model updating moves, as follows. We begin by considering the move which adds a new log-linear parameter to the model. In order to perform this move, we need both to identify and to count the number of models which can be obtained by adding a single parameter to the present model. Since we are assuming that all of the main source terms are present in the model, we have that any second-order interaction term can always be added if it is absent from the present model. We also recall that, since we restrict our attention to hierarchical models, for any lth-order interaction term to be added, there must be at least l $(l-1)$th-order interaction terms in the present model. For model m, and starting with the highest-order interaction terms, we successively check to see whether or not we can add an lth order term by checking to see if there are enough $(l-1)$th-order terms. If there are, then we identify which, if any, lth-order terms it is possible to add, discounting those already present, and add these to a permissibility set, P_m^+. We then move to the $(l-1)$th-order terms until we reach the second-order interaction terms all of which are included in P_m^+ except those already present in the current model. We then propose to add a parameter chosen at random from P_m^+.

This process is simplified in the code by the ordering of the log-linear terms presented

above. To verify whether or not the parameter $\theta^c \equiv \theta_j$, where

$$j = \sum_{i:S_i \in c} 2^{i-1},$$

can be added to the model we need only check for the presence of the parameters θ_l such that $l = (\sum_{i:S_i \in c} 2^{i-1}) - 2^{t-1}$, for all t such that $S_t \in c$. For example, consider the parameter $\theta_7 \equiv \theta^{ABC}$. If this parameter is to be added, the model needs to contain the parameters $\theta_3 \equiv \theta^{AB}$, $\theta_5 \equiv \theta^{AC}$ and $\theta_6 \equiv \theta^{BC}$. This simple sequential approach provides a computationally very efficient algorithm for generating the set P_m^+.

In order to delete a parameter from the current model, we proceed similarly, by examining each parameter in the model and adding it to the permissibility set for deletion, P_m^-, if no other parameter in the model requires it to be present. Clearly, this permissiblity set is the set consisting of maximal elements of the model, which we use to represent the model for brevity of notation. Once we have the set P_m^-, we randomly select a parameter for deletion from that set, each with probability $1/|P_m^{-1}|$.

Having selected a parameter to add, for example, we must next propose a value for that parameter in the new model. We take the proposal distribution to be $N(\mu, \sigma^2)$, where either both μ and σ^2 are fixed, the fixed proposal, or we use pilot tuning by running the Markov chain Monte Carlo simulation for the saturated model in order to obtain a mean and variance for each log-linear parameter, the saturated proposal. Generally, for the fixed proposal, we set $\mu = 0$, since we have no prior knowledge of the interactions between sources. It would seem intuitively plausible that the proposal density using pilot tuning would be better, in the sense that the acceptance probability of the proposed jump would be higher. This is because we are using information about the log-linear parameter being proposed within the reversible jump step. This appeared to be the case with all of the datasets considered, with the acceptance probabilities of the proposals being greater than if the parameters are fixed. For example, for the injury dataset studied in § 5, the acceptance probabilities for a move are approximately 25% for the saturated proposal and 20% for the fixed proposal using $\mu = 0$ and $\sigma^2 = 1$.

Having selected a move type and then values for any new parameters, we perform the accept/reject step. Suppose that we are in model m, that we have independent priors $p_c(\theta^c)$ for all $c \in \mathscr{P}(S)$ and we propose to add a new parameter θ^c, for $c \in P_m^+$, with a proposed value $u \sim q_c(u)$. This move is then accepted with probability $\min(1, A)$, where

$$A = \frac{L_{m \cup c}(p; n) p_c(u) |P_{m \cup c}^-|^{-1}}{L_m(p; n) q_c(u) |P_m^+|^{-1}}$$

with the Jacobian in (3) equal to unity. The acceptance probabilities for deletion moves are similarly defined using the reciprocal relationship in (4).

An obvious extension to the move-types proposed above is to consider the addition/deletion of more than one parameter at a time. However, the acceptance rate of such moves decreases rapidly as the number of parameters involved increases. In such a case, it is possible to consider the multivariate normal distribution as the proposal distribution, with a general covariance matrix, so that the log-linear parameters are not assumed to be independent. The covariance between each pair of the log-linear parameters can once more be obtained by considering the saturated model and using pilot tuning via a Markov chain Monte Carlo simulation.

An alternative approach is to consider switching the labels of two or more elements of S. For example, in the four-source case consider the model $m = \{AB, BCD\}$. Two elements

of S are selected uniformly, A and C, say. We thus propose to move from model m to model $m' = \{BC, ABD\}$ obtained by simply swapping occurrences of A in m by C and vice versa. The new model parameters that are 'born' have values drawn from the saturated proposal described above. Larger moves of this type have considerably lower acceptance probabilities, but their inclusion does appear to improve the overall mixing of the chain, especially for datasets with large numbers of sources and certain posterior model probability structures. Thus, since it necessitates only minor additional computational expense, the inclusion of these moves may be generally beneficial.

5. THE INJURY DATA EXAMPLE AND COST EFFICIENCY

5·1. *Background*

Yip et al. (1995) provide data on the number of injuries in adolescents, for a school district in Allegheny County, Pennsylvania, using observed absences, A; medical records, B; a monthly survey, C; and a four-month survey, D. The first two forms of data were taken from school records and the data collection was performed between 1st September and 31st December 1991. The background to the data collection and detailed descriptions of the sources are given by LaPorte et al. (1995). The data were collected following an earlier study which concluded, by matching with medical records, that three-month surveys were not particularly effective methods for accurate counting of injuries. This second study was intended to be more reliable and included, as additional sources, school attendance records and two forms of regular survey in which individuals were asked to recall any injuries that they had sustained during the relevant period. The data are summarised in Table 1.

Table 1. *Adolescent injuries in a school district in Allegheny County, Pennsylvania, 1 September–31 December* 1991

		$D=1$		$D=0$	
		$C=1$	$C=0$	$C=1$	$C=0$
$A=1$	$B=1$	11	3	2	0
$A=1$	$B=0$	7	3	3	4
$A=0$	$B=1$	13	1	25	3
$A=0$	$B=0$	31	3	35	?

Sources: A, attendance records; B, medical records; C, one-month recall; D, four-month recall.

5·2. *Priors*

As discussed above, a previous study suggested that irregular surveys did not tend to be very reliable. However, it was expected that the one-month surveys, together with attendance and medical records, would provide considerably more accurate counts of the true number of injuries. Thus, we might want to adopt a prior for N which was centred on a value close to the observed number of individuals since we might expect close to 100% ascertainment. However, we cannot be completely sure that sources C and D are as accurate as we would like, so this might be reflected by a fairly large prior variance. Thus, we might choose a gamma hyperprior for λ with a mean and standard deviation equal to 144, the observed number of individuals. Of course, there are many other similarly

vague prior structures which could be adopted. We conducted a sensitivity study by adopting a wide variety of priors which included varying the gamma hyperprior parameters and adopting the Jeffreys prior for N. However, the study suggested that posterior inference was broadly insensitive to the choice of prior for N in this case, most likely because of the strength of the data.

We have no substantive prior information concerning the correlation between sources. Thus, we take the log-linear parameters to be normally distributed about zero, so that there is no prior preference for the type of correlation between any of the sources, with an unknown variance assumed to have an inverse gamma hyperprior, giving prior mean and variance of 1. This seemed a sensible value since previous analyses of similar data rarely showed θ values beyond the range $[-2, 2]$. Thus, we have

$$\theta|\sigma^2 \sim N(0, \sigma^2), \quad \sigma^2 \sim \Gamma^{-1}(3, 2).$$

It could be argued that we might expect the correlation, if any, between sources C and D to be positive, i.e. students tended to be similarly diligent in their approach to the two surveys. This could be incorporated into the analysis by adopting a normal prior distribution with positive mean. The variance of the prior would reflect the strength of this belief. Once again, we observe that the posterior inference obtained is fairly robust against any such changes, and particularly robust with respect to the parameter of primary interest, namely the posterior estimate of N.

Finally, we adopt a flat prior over model space, with no explicit penalty for models with large numbers of parameters. In the absence of any a priori information, this seemed the most sensible approach and, once again, posterior inference seems fairly robust against changes in the prior distribution for models.

5·3. *Simulation and results*

Standard diagnostic techniques (Brooks & Giudici, 2000; Brooks & Roberts, 1998) show that the Markov chain appears to converge well within one million iterations and to be mixing well. Since computing time for runs of this magnitude are fairly small, we run the Markov chain for a total of two million iterations, basing inference upon the second half of the observations obtained. We also repeated the simulation for a variety of priors as discussed above.

During the simulation, a total of 109 different models are observed, i.e. proposed and accepted, within the Markov chain Monte Carlo procedure, out of a total of 168 possible models. The simulations were repeated a number of times and the posterior model probability estimates did not vary within the first three decimal places, so that we can be reasonably sure that the Monte Carlo errors for simulations of this length are negligible and that models not visited at all do indeed have zero posterior mass to three decimal places.

The Bayesian analysis is summarised in Table 2, which provides the posterior model probabilities for the a posteriori most probable models, together with the corresponding posterior means, modes and 95% highest posterior density intervals for N. The expected mean of the precision, σ^{-2}, under the posterior distribution is 2·52, compared to the prior mean of 1·5.

Yip et al. (1995) do not consider the model selection problem, so, for comparison, Table 2 provides the posterior expected AIC and BIC, together with the DIC suggested in an unpublished report by D. J. Spiegelhalter, N. G. Best and B. P. Carlin. Other model selection procedures do exist; for example Fienberg et al. (1999) consider model selection

Table 2. *The posterior model probabilities for the models with probability greater than 0·05, together with posterior mean, mode and 95% highest posterior density interval,* HPDI, *and information criteria values*

Model	Posterior probability	Posterior mean	Posterior mode	95% HPDI	Expected AIC	Expected BIC	DIC
$\{AC, AD, B\}$	0·280	148	147	(144, 153)	77·62	81·87	72·29
$\{AC, AD, CD, B\}$	0·099	150	148	(145, 158)	79·64	84·60	73·01
$\{AB, AC, AD\}$	0·093	148	147	(144, 154)	78·95	83·91	72·45
$\{AC, BC, AD\}$	0·060	148	147	(144, 155)	80·27	85·22	73·74
Model-averaged		150	147	(144, 158)			

A, attendance records; *B*, medical records; *C*, one-month recall records; *D*, four-month recall records. DIC, criterion suggested by D. J. Spiegelhalter, N. G. Best and B. P. Carlin.

procedures without quantitatively incorporating model uncertainty into their estimates. We see from Table 2 that the model with two-way interactions between the attendance records and the one- and four-month recall records, model $\{AC, AD, B\}$, has the highest posterior model probability, which is three times that of the second most probable model. This suggests that there is some degree of correlation between source A and sources C and D. In fact these two interaction terms are present in all four of the 'top' models.

It is worth noting that results very similar to those obtained in Table 2 are observed if we adopt the Jeffreys prior discussed in § 3. The posterior model probabilities for the top four models retain the same ordering with values of 0·272, 0·094, 0·091 and 0·059 respectively. The posterior means and modes remain unchanged. Thus, the chosen priors were indeed reasonably vague and we can be confident that our resulting analysis is strongly data- rather than prior-driven.

We can augment the summary analysis presented in Table 2 and formally quantify the probability associated with any individual log-linear term being present in the model, a technique that is not possible within the classical framework. The posterior probability of the presence of a log-linear term is just the proportion of times that the state of the Markov chain is in any model containing the given term. Table 3 gives the posterior probability of each of the possible individual log-linear interaction terms contained within the models. The strong dependence of the posterior distribution on the two interaction terms AC and AD within the models is clearly shown, with both of the terms having a posterior probability of presence in excess of 0·9.

Table 3. *The posterior probabilities for log-linear interaction terms within the model*

Model parameter	Posterior probability	Model parameter	Posterior probability
AB	0·358	BD	0·195
AC	0·901	ABD	0·077
BC	0·262	CD	0·335
ABC	0·030	ACD	0·060
AD	0·967	BCD	0·014

Table 2 also illustrates the strength of agreement between the information criteria and the posterior model probabilities in terms of which model might be best used for inference. An advantage of our approach is that we can average over all of the models to obtain a single estimate. Of course, this is not the only approach to model averaging. For example

Pauler (1998) shows how the BIC can be used for model averaging in the context of linear models, but similar approaches may be extended to the present context. We can see from Table 2 that the model-averaged posterior mean for the total population size is 150, suggesting that only six injuries were missed by all four sources. The model-averaged 95% highest posterior density interval is (144, 158), at least as wide as that under any single model, properly reflecting our uncertainty as to the 'true' model. Of course in this example, where the models largely predict similar population sizes, the advantage of model averaging over simply picking the 'best' model is not fully demonstrated.

In practice it is common to find several models that are identified as plausible yet give vastly different estimates of the total population size. To illustrate this point we consider the dataset presented by Frischer et al. (1993) concerning the number of drug addicts within the city of Glasgow. For brevity we do not present the actual data or details of the analysis performed, but only give the results obtained using the same approach as used above with a similarly vague prior, and compare them with the corresponding classical analysis. In this example, the three most plausible models within our Bayesian analysis account for over 77% of the total posterior model probabilities. For the second and third most probable models a posteriori, with probabilities 0·171 and 0·151, the total estimated population size ranges from 7694 to 8876. The most probable model has a posterior model probability of 0·455, with corresponding posterior mean of 8119 and standard deviation of 371. The model-averaged mean of N is 8219 with a standard deviation of 636, the latter being considerably larger than that for any single model. The model selected under the corresponding classical analysis was ranked fourth within the Bayesian analysis with posterior model probability of 0·047 and a maximum likelihood estimate of 8494 for N.

The model averaging approach using reversible jump Markov chain Monte Carlo improves upon competing methods by providing a very efficient mechanism for incorporating model uncertainty into the analysis. This removes the need for large numbers of simulations, one for each model, which may prove prohibitive in many applications, to obtain these probabilities via the BIC method discussed above, for example. Without posterior model probability estimates, however obtained, the incorporation of uncertainty due to the model is at best subjective. The advantage of the model averaging approach is that we obtain an objective and quantitative estimate of the uncertainty due to the models. Perhaps most importantly, the model averaging approach tells us when the additional uncertainty attributable to model choice is large compared to the uncertainty within models. This was not the case for the Yip et al. (1995) data but, in the case of the Frischer et al. (1993) data, model uncertainty accounts for around half of the uncertainty about the total population size. This suggests that any predictive inference would seriously underestimate the variability of the population if only a single model were used as often happens in classical analyses.

The posterior capture probability of each source is provided in Table 4, along with the effort, in hours, needed to obtain the data from each source. These posterior capture probabilities are obtained by averaging over the proportion of injuries that appear upon each list in relation to the estimated total number during each iteration of the Markov chain. As we might have expected, the capture rate for source D, the four-month recall, is somewhat smaller than that for C, with the one-month recall survey identifying nearly 85% of all injuries. This seems plausible, since we might expect that recollection of the number of injuries over the past four months might be somewhat less reliable than that over a one-month period.

LaPorte et al. (1995) also perform a classical analysis of the data, using a maximum of

Table 4. *The effort and posterior probabilities of the sources*

Source	Effort (hours)	No. of injuries identified	Posterior capture probability	95% HPDI
A	48	33	0·220	(0·209, 0·229)
B	30	58	0·387	(0·367, 0·403)
C	288	127	0·847	(0·804, 0·882)
D	72	72	0·480	(0·456, 0·500)

A, attendance records; *B*, medical records; *C*, one-month recall; *D*, four-month recall.
HPDI, highest posterior density interval.

three sources, and report the corresponding log-linear model that is deemed to be the best fit to each combination of data sources. For the partial datasets considered by LaPorte et al. (1995), we perform the corresponding Bayesian analysis using the framework detailed above and the priors as specified for the full dataset, for comparison. When considering all of the sources, excluding the four-month student recall, LaPorte et al. (1995) suggest that the model deemed to fit best is that with the single two-way interaction between the attendance and the one-month survey records, giving a maximum likelihood estimate for the total number of injuries of 147 with a corresponding standard error of 3·3. This model is calculated to be the most probable a posteriori under the Bayesian analysis, with a posterior model probability of 0·38. The corresponding posterior mean, mode and standard deviation are 148, 147 and 4·3, respectively.

We observe that the maximum likelihood estimate and posterior mode coincide, but that the classical standard error is somewhat smaller than the corresponding posterior standard deviation. This is almost certainly because of a breakdown of the classical assumption of asymptotic normality. Under this assumption, and under a flat prior as we have here, the posterior mode should be identical to the posterior marginal mean in which case the posterior mode and mean for N should coincide. This suggests that the assumption of asymptotic normality does not quite hold and may well explain the consequential underestimation of the uncertainty associated with the classical estimate of the total population size. Correspondingly, the posterior model-averaged estimate of the total number of injuries, using these sources, is 160, with a standard deviation of 20·9. The higher variance here can be attributed to the explicit inclusion of the uncertainty associated with not only the value of N but also the model itself.

In addition to the analyses performed above, it may also be useful to consider other questions of interest to epidemiologists. One question often asked is how effective are the different sources at capturing individuals. From Table 4, we can see that there appears to be some degree of correlation between expense and posterior probability of capture. However, it is often of interest to quantify the utility of any particular data source with regard to the additional information the source provides, and to offset that against its cost of collection. For example, suppose that we assume that the population does not change over time. Then, within the context of the injury data, is it worthwhile in future to collect the information from sources *A*, the attendance records, and/or *B*, the medical records, since they only capture approximately 22% and 38% of injuries, while other sources are more effective at identifying injuries but involve greater effort? Hay (1997)

discusses the dangers of not using all available sources, particularly when there are a few of them. In § 6 we consider this critical problem in further detail.

6. Cost effectiveness of different sources

6·1. *Ad hoc approaches*

Here, we address the problem of choosing the most cost-effective collection of sources, that give an 'adequate' estimate of the total population size. An ad hoc procedure for choosing the most cost-effective model is undertaken by LaPorte et al. (1995), where the total effort of each combination of sources was plotted against the accuracy of the corresponding inference, given by the coefficient of variation, defined as follows:

$$\text{coefficient of variation} = 100 \times \frac{\text{standard deviation}}{\text{estimated number of injuries}}.$$

The most cost-effective model is then chosen subjectively and is often influenced by either effort or accuracy constraints. Data collection from different sources invariably requires a cost of some sort. This can sometimes be limited, placing a restriction upon the combination of sources that can be used. In such cases, the most accurate combination of sources is required that falls within these constraints. The inverse problem can also be of interest, in that a combination of data sources is required to ensure that the resulting inference has a certain precision. In general, there is a trade-off between the accuracy of the inference and the effort taken in collecting the data.

Such problems are ideally suited to the formal decision-theoretic framework of the Bayesian paradigm. Before we discuss this approach in detail, we first consider an extension to the ad hoc procedure proposed by LaPorte et al. (1995). We consider a graphical Bayesian approach in which model averaging with the full set of data is used to obtain the 'best' estimate for N together with a 95% credible interval. This is compared with the corresponding model-averaged estimates from subsets of the original sources. The corresponding cost and accuracy for each of the submodels, using the different combinations, are illustrated in Fig. 1.

Figure 1 suggests that the subsets of the data with a cost of less than 150 tend to have wide credible intervals. These subsets exclude the one-month recall data, which takes more effort than the others but also identifies the most injuries. Note that, when sources A and B, and A and D, labelled 1 and 3 respectively are used, the corresponding posterior mean of N is less than the total number of identified under all of the sources. In fact, the total observed number of injuries from all data sources does not even lie within the corresponding 95% credible interval when using only sources A and D.

An ad hoc interpretation of Fig. 1 can be used in order to choose the most cost-effective set of sources. For example, we might choose subset 5, which only uses the medical records and one-month recall sources, to be the most cost-effective, with the fifth smallest effort needed and a relatively short 95% credible interval of observations. Alternatively, we might choose subset 6, with is both 'cheap' and provides an accurate point estimate, though the uncertainty is considerably larger than that for subset 5. As discussed earlier, Hay (1997) questions the accuracy of estimates on the basis of subsets of the data, particularly where there are only a small number of sources. It should also be pointed out that analyses of this sort are necessarily retrospective in that they address the question of which set of sources would have been most efficient for this particular dataset. However, the real question of interest is what might be done in the future to ensure that further data are

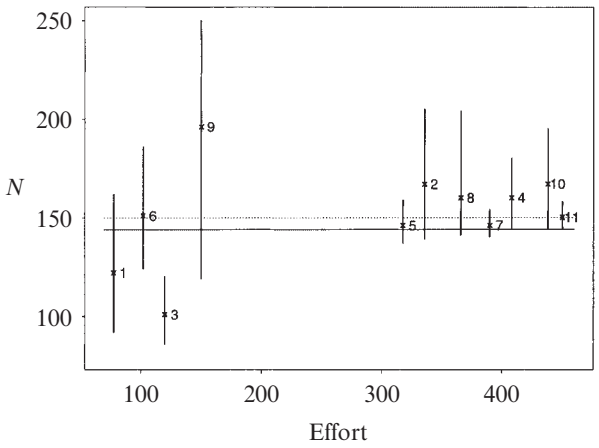

Fig. 1. Estimated values of the total number of adolescent injuries using a Bayesian approach for subsets of the original data, where the vertical lines represent the model-averaged 95% highest posterior density interval and a cross represents the estimated mean of the posterior distribution. The solid horizontal line represents the total number of injuries observed under all of the sources, while the dotted horizontal line indicates the model-averaged estimate of the total number of injuries using all available data and thus represents our 'best guess' at the total population size. The sources used in each model are labelled as: 1, *A* and *B*; 2, *A* and *C*; 3, *A* and *D*; 4, *A* and $(C + D)$; 5, *B* and *C*; 6, *B* and *D*; 7, *B* and $(C + D)$; 8, *A*, *B* and *C*; 9, *A*, *B* and *D*; 10, *A*, *B* and $(C + D)$; 11, *A*, *B*, *C* and *D*; note that the effort for this subset is 438, the same as for subset 10, and hence indistinguishable if plotted at its correct effort. Note: $(C + D)$ refers to the combination of sources *C* and *D*, the recall records, so that the individuals appearing from this source are present in either *C*, *D* or both, while an individual not listed on this source did not appear in either of sources *C* and *D*.

collected most efficiently. To address this question, removing the associated subjectivity of the ad hoc approaches described above, we adopt a quantitative Bayesian decision-theoretic approach by constructing a loss function which combines losses in accuracy and the associated costs of data collection.

6·2. *Loss functions*

In the case of adolescent injuries, the loss function needs to take into account both the accuracy of the Bayes estimate and time expended in collecting the data from the sources in each of the data subsets under consideration. The loss function is also dependent upon the unknown total number of adolescent injuries *N*. Madigan & York (1997) refer to the relative squared error loss function, which can be expressed as

$$L(N_s, N) = \frac{(N - N_s)^2}{N^2},$$

where N_s is the model-averaged estimate of *N* using data sources $s \subseteq S$. This loss function essentially weights estimation errors by the true value, so that errors are measured on a

relative rather than absolute scale. The Bayes estimate associated with this loss function is of closed form, and can be easily shown to be

$$\hat{N}_s = \frac{E_\pi(N^{-1}|s)}{E_\pi(N^{-2}|s)}.$$

However, the cost of the associated data collection also needs to be included in the loss function, so we consider modifying the relative squared error loss function by incorporating a function for the cost, C_s, of the hours involved in collecting the data from all the sources in s.

The overall loss function can then be considered to be a function of both the loss associated with the accuracy of the estimate of N and the cost of the corresponding data. The loss function may depend upon both the accuracy required and the cost limitations. For illustration, we consider a general loss function of the form

$$\begin{aligned} L(N_s, N, C_s) &= v f_1\{L(N_s, N)\} + (1-v) f_2(C_s) \quad (v \in [0,1]) \\ &= v f_1\left\{\frac{(N-N_s)^2}{N^2}\right\} + (1-v) f_2(C_s), \end{aligned}$$

for constant v and functions f_1 and f_2. The constant v represents the relative importance of the error of the estimate and the corresponding cost incurred. The most accurate model, disregarding cost, can then be calculated by setting $v = 0$ in the overall loss function. Similarly, when the accuracy of the estimate is irrelevant compared to the effort needed for the collection of the data, we use $v = 1$.

This general expression for the overall loss function permits many possible forms combining the error of the estimate and the cost. It would seem sensible to consider an additive or multiplicative loss function, since they are readily interpretable. We might define an additive loss function by defining f_1 to be the identity function and f_2 to be a relative function, that is $f_2(z) = z/T$, where T is a constant. In this example, it seems reasonable to set T to be the total effort in collecting data from all of the sources available. The loss function can therefore be expressed as follows:

$$L(N_s, N, C_s) = v\left\{\frac{(N-N_s)^2}{N^2}\right\} + (1-v)\left(\frac{C_s}{T}\right). \tag{5}$$

Similarly, for a multiplicative loss function we might define both f_1 and f_2 to be the log function. Hence the loss function can be written as

$$L(N_s, N, C_s) = v \log\left\{\frac{(N-N_s)^2}{N^2}\right\} + (1-v)\log(C_s).$$

In practice the loss functions should be derived in tandem with the prior elicitation process so that expert opinion is incorporated into both. Once the functions f_1 and f_2 have been defined, the set of sources considered 'optimal' under the corresponding loss function is the one that minimises the expected loss under the posterior distribution, i.e. the Bayes risk, and so minimises

$$\begin{aligned} E\{L(N_s, N, C_s)\} &= E\left[v f_1\left\{\frac{(N-N_s)^2}{N^2}\right\} + (1-v) f_2(C_s)\right] \\ &= vE\left[f_1\left\{\frac{(N-N_s)^2}{N^2}\right\}\right] + (1-v) f_2(C_s), \end{aligned}$$

since $f_2(C_s)$ is constant, given s, and v is a constant defining the relative importance between the cost and accuracy of the resulting inference. This expectation can be calculated through a Markov chain Monte Carlo procedure, estimating the expected loss under the posterior predictive distribution.

6·3. *Simulation*

We begin by assuming that the population is time-invariant and proceed by drawing a sample of model parameters from the posterior distribution associated with the original data. We then generate possible future datasets from the model using this sample of model parameters. Having obtained these new datasets, we seek to find the combination of data sources that minimises our expected loss, taken with respect to these datasets and their corresponding posterior distributions for the model parameters.

We begin with an initial simulation sampling from the posterior distribution corresponding to the original data. We subsample the final 500 000 iterations taking every 500th to obtain 1000 roughly independent draws from the posterior. From each of these we simulate a new dataset from the multinomial model with parameters corresponding to that particular realisation from the original chain. Then, for each dataset we perform a cost-efficiency analysis to obtain the model-averaged most cost-effective set of sources for any given v.

This procedure can be extremely computationally intensive, since separate Markov chain Monte Carlo runs need to be performed for each separate set of sources and for each simulated dataset. However, the procedure can be automated and run in the background over several days if necessary.

6·4. *Cost-efficient data sources*

Obviously, the choice of loss function should be made within the context of the problem at hand. To illustrate our approach, we consider the additive loss function of (5) for the injury data. We use the Markov chain Monte Carlo procedure outlined above to obtain samples from the relevant posterior distributions, for each of the simulated datasets generated. The combinations of sources which minimise this loss function can be summarised as a function of v, as follows: sources A and B, subset 1, for $0 \leqslant v \leqslant 0{\cdot}5424$; sources B and D, subset 6, for $0{\cdot}5424 \leqslant v \leqslant 0{\cdot}9453$; sources B and C, subset 5, for $0{\cdot}9453 \leqslant v \leqslant 0{\cdot}9830$; and sources B and $(C + D)$, subset 7, for $0{\cdot}9830 \leqslant v \leqslant 1$. The remaining data-subsets are inadmissible under this loss function. Thus, given a value of v, perhaps by a relevant expert, the most cost-effective set of sources is easily identified.

Recall that large values of v place most importance on accuracy, whilst small values place importance on low cost. Thus, we can see that for values of v between 0 and 0·54 we prefer the cheapest option, whilst for values of v over 0·98, we prefer the most accurate. The intuitive plausibility of these possibilities as providing a compromise between accuracy and cost is further demonstrated by Fig. 1, where an ad hoc interpretation may have selected the same four data subsets as being 'optimal'. However, this decision-theoretic approach has the advantage that it provides a formal framework for quantifying the trade-off between data collection costs and inferential accuracy, though it involves some additional computational expense.

In practice, this framework may be extremely useful for deciding whether or not to continue collecting data from any particular source, so long as it is reasonable to assume that the performance of any source remains unchanged in the future. If this were not the

case, then the analysis could be redone with respect to a predictive distribution for the future population, though this would require rather subjective assumptions in general.

7. Discussion

Though we have only discussed a simple 2^4 example here, the methodology works equally well with much larger numbers of sources. Though the number of distinct models rises exponentially with the number of sources, computational expense remains reasonable even for high-dimensional problems since it is only the most likely models which will be explored. Though we might expect problems with mixing for very high-dimensional problems, greater than 10, say, and where many models with high-order interactions have high posterior probability, we have not experienced any problems ourselves and would expect that the inclusion of additional 'large' moves could minimise any such problems, should they arise.

We might also consider extending this work to more general problems of missing data in contingency tables. Certainly the extension to cases where more than a single cell is unobserved is straightforward, though the extension beyond the $2^{|S|}$ case, where factors have more than two levels, begins to increase the computational expense of keeping track of the lexicographic ordering for the parameters. Such problems might arise where individuals were not simply seen, but observed to be in one of several states. For very large-dimensional tables with high numbers of classes, it is likely that other approaches would be required.

Alternative methods to log-linear models that have been recently applied to the problem of census undercount for $2^{|S|}$ contingency tables include logistic regression models (Bray & Wright, 1998), association models (Becker & Clogg, 1989) and Rasch and latent class models (Coull & Agresti, 1999; Fienberg et al., 1999). In particular, the latter provide an alternative to the assumption of there being no $|S|$-way interaction in the model. There is no real reason to reject any of these models a priori and so it would be interesting to consider a more general class of models encompassing all of these when tackling the model selection problem. This would provide an extremely powerful set of tools for understanding and predicting these populations and is the focus of current research.

Acknowledgement

The authors gratefully acknowledge the help and advice provided by Jon Forster, Ernie Hook, Ron Regal, two anonymous referees and the editor. We also acknowledge the financial assistance of the UK Engineering and Physical Sciences Research Council in funding the research of both authors.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Becker, M. & Clogg, C. (1989). Analysis of sets of two-way contingency tables using associated models. *J. Am. Statist. Assoc.* **84**, 142–51.

Bray, I. & Wright, D. E. (1998). Application of Markov chain Monte Carlo methods to modelling birth prevalence of Down syndrome. *Appl. Statist.* **47**, 589–602.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *Statistician* **47**, 69–100.

Brooks, S. P. & Giudici, P. (2000). MCMC convergence assessment via two-way ANOVA. *J. Comp. Graph. Statist.* **9**, 266–85.

Brooks, S. P. & Roberts, G. O. (1998). Diagnosing convergence of Markov chain Monte Carlo algorithms. *Statist. Comp.* **8**, 319–35.

COULL, B. & AGRESTI, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* **55**, 294–301.

DAWID, A. P. & LAURITZEN, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

DELLAPORTAS, P. & FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**, 615–33.

EDWARDS, D. & HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72**, 339–51.

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**, 591–603.

FIENBERG, S. E., JOHNSON, M. S. & JUNKER, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. R. Statist. Soc.* A **162**, 383–405.

FRISCHER, M., LEYLAND, A., CORMACK, R., GOLDBERG, D. J., BLOOR, M., GREEN, S. T., TAYLOR, A., COVELL, R., MCKEGANEY, N. & PLATT, S. (1993). Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in Glasgow, Scotland. *Am. J. Epidem.* **138**, 170–81.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

HAY, G. (1997). The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* **44**, 515–20.

HOOK, E. B., ALBRIGHT, S. G. & CROSS, P. K. (1980). Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York State. *Am. J. Epidem.* **112**, 750–8.

HOOK, E. B. & REGAL, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidem. Rev.* **17**, 243–64.

KING, R. & BROOKS, S. P. (2001). Prior induction in log-linear models for general contingency table analysis. *Ann. Statist.* To appear.

LAPORTE, R. R., DEARWATER, S. R., CHANG, Y., SONGER, T. J., AARON, D. J., ANDERSON, R. L. & OLSEN, T. (1995). Efficiency and accuracy of disease monitoring systems: Application of capture-recapture methods to injury monitoring. *Am. J. Epidem.* **142**, 1069–77.

MADIGAN, D. & YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84**, 19–31.

O'HAGAN, A. (1998). Eliciting expert beliefs in substantial practical applications. *Statistician* **47**, 21–36.

PAULER, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika* **85**, 13–27.

RICHARDSON, S. & GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

YIP, P. S. F., BRUNO, G., TAJIMA, N., SEBER, G. A. F., UNWIN, N., CHANG, Y., FIENBERG, S. E., JUNKER, B. W., LAPORTE, R. E., LIBMAN, I. M. & MCCARTY, D. J. (1995). Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *Am. J. Epidem.* **142**, 1059–68.

[*Received March* 2000. *Revised December* 2000]