Journal of Statistical Software

June 2014, Volume 58, Issue 7. http://www.jstatsoft.org/

conting: An R Package for Bayesian Analysis of Complete and Incomplete Contingency Tables

Antony M. Overstall
University of St Andrews

Ruth King
University of St Andrews

Abstract

The aim of this paper is to demonstrate the R package **conting** for the Bayesian analysis of complete and incomplete contingency tables using hierarchical log-linear models. This package allows a user to identify interactions between categorical factors (via complete contingency tables) and to estimate closed population sizes using capture-recapture studies (via incomplete contingency tables). The models are fitted using Markov chain Monte Carlo methods. In particular, implementations of the Metropolis-Hastings and reversible jump algorithms appropriate for log-linear models are employed. The **conting** package is demonstrated on four real examples.

Keywords: contingency tables, capture-recapture studies, reversible jump, log-linear models.

1. Introduction

Contingency tables (see, e.g., Agresti 2007) are formed when a population is cross-classified according to a series of categories (or factors). Each cell count of the contingency table gives the number of units observed under a particular cross-classification. The purpose of forming a contingency table is to identify the dependence structure between the factors, i.e., to identify associations (or interactions) between the factors, using statistical models. Additionally, incomplete contingency tables formed from capture-recapture studies can be used to estimate closed populations (Fienberg 1972). Here some of the factors correspond to sources which have or have not observed members of a target population. Cell counts corresponding to not being observed by any of the sources are missing (or unknown). However they can be estimated by fitting a statistical model to the observed cell counts and predicting the missing cell counts. Note that our definition of an incomplete contingency table differs from that of, e.g., Mantel (1970) who defines incomplete contingency tables to be those with structural zeros. We can also distinguish our concept of incomplete contingency tables from those where

the cell counts are misclassified or are only partially observed due to non-response from units of the population, see, e.g., Gelman, Carlin, Stern, and Rubin (2004, Sections 21.5 and 21.6) and Tan, Tian, and Ng (2010, Chapter 4).

Log-linear models are typically used to model the observed cell counts, where the log of the expected cell count is proportional to a linear combination involving unknown model parameters. Each interaction term between two or more factors is associated with a set of model parameters and if these model parameters are non-zero it indicates that there exists an association between these factors. Every unique combination of interactions defines its own log-linear model. Therefore identifying interactions is a model determination problem.

Log-linear models are a special case of generalized linear models (GLMs; e.g., McCullagh and Nelder 1989). Therefore, within R (R Core Team 2014), log-linear models can be fitted via classical maximum likelihood using the `glm()` function. Model determination can then be achieved using, for example, differences in deviance or the Akaike information criterion (AIC). Still under the classical approach, the `loglm()` function in the **MASS** package (Venables and Ripley 2002) provides an alternative approach to fitting log-linear models. Note that `loglm()` is a user-friendly wrapper to the `loglin()` function in the **stats** package (R Core Team 2014).

In this paper and in **conting** (Overstall 2014), the Bayesian approach is used. For a review of the Bayesian approach to contingency tables with log-linear models, see, e.g., Forster (2010). The Bayesian approach involves evaluating the posterior distribution of model parameters and model indicators. Typically, the posterior is analytically intractable and requires approximation. The standard approach is to use Markov chain Monte Carlo (MCMC) methods to generate a sample from the posterior.

The R package **conting** facilitates the Bayesian analysis of complete and incomplete contingency tables using log-linear models. This is accomplished with the use of MCMC methods that are particularly suitable for log-linear models. This paper demonstrates **conting** and is structured as follows. In Section 2 we describe log-linear models for the analysis of contingency tables. In Section 3 we describe suitable MCMC computational algorithms that are required to generate posterior samples. We also describe how these methods are implemented in **conting**. We conclude with an extensive examples section (Section 4) that fully exhibits the capabilities of **conting** for complete and incomplete contingency tables. Note that **conting** is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=conting`.

2. Log-linear models

In this section we set-up our notation and define the concept of a log-linear model under a Bayesian approach in the presence of model uncertainty. This set-up closely follows that of Overstall and King (2014).

Suppose there are a total of F factors (and/or sources) such that the jth factor, for $j = 1, \ldots, F$, has l_j levels. The corresponding contingency table has $n = \prod_{j=1}^{F} l_j$ cells. Let $\mathbf{y}$ be the $n \times 1$ vector of cell counts where each element of $\mathbf{y}$ is denoted by $y_{\mathbf{i}}$ with $\mathbf{i} = (i_1, \ldots, i_F)$ identifying the combination of factor levels that cross-classify this cell. Let $\mathcal{S}$ denote the set of all n cross-classifications and thus the set of all cells in the table. Let $N = \sum_{\mathbf{i} \in \mathcal{S}} y_{\mathbf{i}}$ denote the total population size which can, in the case of incomplete contingency tables, be unknown, due to certain elements of $\mathbf{y}$ being unobserved.

For example, consider the alcohol, obesity and hypertension (AOH) data given by Knuiman and Speed (1988) and included as an example dataset in **conting**. Here, 491 people from Western Australia are cross-classified according to the following $F = 3$ factors: alcohol intake ($l_1 = 4$ levels); obesity ($l_2 = 3$ levels) and hypertension ($l_3 = 2$ levels). Therefore, there are $n = l_1 \times l_2 \times l_3 = 24$ cells in the corresponding contingency table and the total population size is $N = 491$. See Section 4.1 for the associated Bayesian analysis of this table using **conting**, which identifies interactions between the factors.

For a log-linear model it is assumed that

$$y_{\mathbf{i}} \sim \text{Poisson}\left(\mu_{\mathbf{i}}\right), \tag{1}$$

independently, for $\mathbf{i} \in \mathcal{S}$, where $\log \mu_{\mathbf{i}}$ is written as a linear combination of the intercept, main effect and interaction parameters (see, e.g., Overstall and King 2014). However for identifiability, these parameters are constrained using, for example, sum-to-zero or corner-point constraints. In this paper and in **conting**, we use sum-to-zero constraints. The log of the expectation, $\mu_{\mathbf{i}}$, of $y_{\mathbf{i}}$ can then be written as

$$\log \mu_{\mathbf{i}} = \eta_{\mathbf{i}} = \phi + \mathbf{x}_{\mathbf{i}}^{\top} \boldsymbol{\theta}, \tag{2}$$

where $\eta_{\mathbf{i}}$ is referred to as the linear predictor, $\phi \in \mathbb{R}$ is the unknown intercept parameter, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the $p \times 1$ vector of unknown regression parameters and $\mathbf{x}_{\mathbf{i}}$ is the $p \times 1$ design vector that identifies which elements of $\boldsymbol{\theta}$ are applicable to cell $\mathbf{i} \in \mathcal{S}$. Let $\boldsymbol{\beta} = \left(\phi, \boldsymbol{\theta}^{\top}\right)^{\top}$ be the $(p+1) \times 1$ vector of log-linear parameters. We can write (2) in matrix form as

$$\log \boldsymbol{\mu} = \boldsymbol{\eta} = \phi \mathbf{1}_n + \mathbf{X}\boldsymbol{\theta} = \left(\mathbf{1}_n, \mathbf{X}\right) \boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ are $n \times 1$ vectors with elements given by $\mu_{\mathbf{i}}$ and $\eta_{\mathbf{i}}$, $\mathbf{1}_n$ is an $n \times 1$ vector of ones, $\mathbf{X}$ is the $n \times p$ design matrix with rows given by $\mathbf{x}_{\mathbf{i}}$ and $\log(\cdot)$ is applied element-wise.

The specification given by (2), and equivalently by (3), assumes that we know which interactions are present in the model. This will not be the case in practice, so we introduce an unknown model indicator, m, which denotes a combination of interactions. We now write the log-linear model as

$$\log \mu_{\mathbf{i}} = \eta_{\mathbf{i}} = \phi + \mathbf{x}_{m,\mathbf{i}}^{\top} \boldsymbol{\theta}_m, \tag{4}$$

where $\boldsymbol{\theta}_m$ is the $p_m \times 1$ vector of regression parameters for model m and $\mathbf{x}_{m,\mathbf{i}}$ is the $p_m \times 1$ design vector that identifies which elements of $\boldsymbol{\theta}_m$ are applicable to cell $\mathbf{i} \in \mathcal{S}$ for model m. Let $\boldsymbol{\beta}_m = \left(\phi, \boldsymbol{\theta}_m^{\top}\right)^{\top}$ denote the vector of log-linear parameters for model m. In matrix form this model is

$$\log \boldsymbol{\mu} = \boldsymbol{\eta} = \phi \mathbf{1}_n + \mathbf{X}_{\mathbf{m}} \boldsymbol{\theta}_m = \left(\mathbf{1}_n, \mathbf{X}_m\right) \boldsymbol{\beta}_m, \tag{5}$$

where $\mathbf{X}_m$ is the $n \times p_m$ matrix with rows given by $\mathbf{x}_{m,\mathbf{i}}$.

Let $\mathcal{M}$ be the set of all models we wish to consider. In this paper and in **conting**, $\mathcal{M}$ consists of hierarchical log-linear models (e.g., Dellaportas and Forster 1999). Note that hierarchical models include, as a subset, the classes of graphical and decomposable models. Hierarchical models adhere to the principle of marginality (e.g., Fox 2002, p. 135), i.e., we cannot have a higher-order interaction unless all the constituent lower-order interactions are included in the model. Typically, see, e.g., Dellaportas and Forster (1999), the simplest (or minimal) model we wish to consider is the so-called independence model, i.e., the model with main effects for

all factors but no interactions. We also now define the concept of the maximal model, i.e., the most complex model we wish to consider. This is usually accomplished by specifying the highest-order interaction terms that we will consider. The saturated model, for a complete contingency table, is the one that contains the F-way interaction between all factors.

To complete the model specification under the Bayesian approach, we specify a joint prior distribution for ϕ, $\boldsymbol{\theta}_m$ and m which is denoted by $\pi(\phi, \boldsymbol{\theta}_m, m)$. We decompose this distribution as follows

$$\pi(\phi, \boldsymbol{\theta}_m, m) = \pi(\phi, \boldsymbol{\theta}_m | m)\pi(m),$$

where $\pi(\phi, \boldsymbol{\theta}_m|m)$ is the joint prior distribution of ϕ and $\boldsymbol{\theta}_m$ conditional on model $m \in \mathcal{M}$ and $\pi(m)$ is the prior model probability of model $m \in \mathcal{M}$, such that $\sum_{m\in\mathcal{M}} \pi(m) = 1$.

In this paper, and in **conting**, we assume a position of having weak prior information and wish to specify prior distributions that reflect this position. However due to Lindley's paradox (see, e.g., O'Hagan and Forster 2004, pp. 77–79) care must be taken when specifying prior distributions that reflect weak prior information since the posterior model probabilities will be sensitive to the scale of the prior variance.

One approach from the literature is to use a "default" prior distribution (Kass and Wasserman 1996). For examples of such prior distributions which are directly applicable to log-linear models, see Dellaportas and Forster (1999), Ntzoufras, Dellaportas, and Forster (2003), Sabanés-Bové and Held (2011) and Overstall and King (2014). We use the generalized hyper-g prior proposed by Sabanés-Bové and Held (2011) where for a log-linear model, we decompose the joint prior distribution of ϕ and $\boldsymbol{\theta}_m$ as

$$\pi(\phi, \boldsymbol{\theta}_m|m) = \pi(\phi)\pi(\boldsymbol{\theta}_m|m),$$

with $\pi(\phi) \propto 1$ and

$$\boldsymbol{\theta}_m|\sigma^2, m \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 n \left(\mathbf{X}_m^\top \mathbf{X}_m\right)^{-1}\right),$$

with $\sigma^2 > 0$ an unknown hyperparameter with a hyper-prior distribution given by $\sigma^2 \sim \mathrm{IG}\left(\frac{a}{2}, \frac{b}{2}\right)$, where a and b are specified hyperparameters. The Sabanés-Bové and Held prior distribution is a generalization to GLMs of the Zellner g-prior (Zellner 1986) for linear models. It can be interpreted as being the posterior distribution from a locally uniform prior and an imaginary sample where $1/\sigma^2$ indicates the size of this "prior sample" (Dellaportas and Forster 1999). If $\sigma^2 = 1$, then the Sabanés-Bové and Held prior distribution provides the same information as a prior sample of one observation and the prior reduces to the unit information prior (Ntzoufras *et al.* 2003).

Both the Sabanés-Bové and Held and unit information priors are implemented in **conting**. The user can specify the value of the hyperparameters a and b, under the Sabanés-Bové and Held prior. They have default values of $a = b = 10^{-3}$ and we use these default values whenever we employ the Sabanés-Bové and Held prior for the examples in Sections 4.3 and 4.4.

For the prior model probabilities we assume a uniform prior over the model space, i.e., $\pi(m) = |\mathcal{M}|^{-1}$ where $|\mathcal{M}|$ denotes the number of models in $\mathcal{M}$.

Note that, as an alternative to the model specification given in (1), we can assume that

$$\mathbf{y}|N, \boldsymbol{\rho} \sim \mathrm{Multinomial}\left(N, \boldsymbol{\rho}\right),$$

where the elements of $\boldsymbol{\rho}$ are given by

$$\rho_{\mathbf{i}} = \frac{\mu_{\mathbf{i}}}{\sum_{\mathbf{i}\in\mathcal{S}} \mu_{\mathbf{i}}}. \tag{6}$$

Under the multinomial specification, the intercept, ϕ, is unnecessary since it will cancel in the numerator and denominator of (6). For complete contingency tables, Forster (2010) shows that the joint posterior distribution of $\boldsymbol{\theta}_m$ and m are identical under both the Poisson and multinomial models, if the joint prior distribution for ϕ and $\boldsymbol{\theta}_m$ is specified as $\pi(\phi, \boldsymbol{\theta}_m|m) = \pi(\boldsymbol{\theta}_m|m)$, i.e., as is true for the Sabanés-Bové and Held and unit information priors.

Overstall, King, Bird, Hutchinson, and Hay (2014) extend this result to incomplete contingency tables and show that the joint posterior distribution of $\boldsymbol{\theta}_m$, m and the missing cell counts are identical under the Poisson and multinomial models if we adopt the same prior structure as above for ϕ and $\boldsymbol{\theta}_m$, and the prior distribution for the unknown total population size, N, under the multinomial model is of the form $\pi(N) \propto N^{-1}$, i.e., the Jeffreys prior (Madigan and York 1997).

It follows that the MCMC methods detailed in Section 3 allow us to evaluate the posterior distribution under either the Poisson or multinomial model formulations.

Under the Sabanés-Bové and Held, and unit information prior distributions described above, the posterior distribution is analytically intractable necessitating the use of the MCMC methods, detailed in Section 3, to estimate, for example, the posterior model probabilities. Under the subset of decomposable models and the multinomial model formulation, Dawid and Lauritzen (1993) show that by using a hyper-Dirichlet prior distribution the posterior model probabilities are available in closed form. Madigan and York (1997) used this result to estimate closed population sizes from incomplete contingency tables. However, as noted by Dellaportas and Forster (1999), we should not restrict ourselves to the less flexible class of decomposable models for solely computational reasons. Hence we consider the much richer class of hierarchical models and use the prior distributions described above.

3. Methods and implementations

3.1. Posterior distributions

In the case of complete contingency tables we evaluate the joint posterior distribution of ϕ, $\boldsymbol{\theta}_m$, σ^2 (if unknown) and m given by

$$\pi(\phi, \boldsymbol{\theta}_m, \sigma^2, m|\mathbf{y}) \propto \pi(\mathbf{y}|\phi, \boldsymbol{\theta}_m, m)\pi(\boldsymbol{\theta}_m|\sigma^2, m)\pi(\sigma^2)\pi(m), \tag{7}$$

where $\pi(\mathbf{y}|\phi, \boldsymbol{\theta}_m, m)$ is the likelihood function under model $m \in \mathcal{M}$. This distribution is analytically intractable so we generate a sample from it using MCMC methods. In particular we use the reversible jump algorithm (Green 1995). We briefly describe a particular implementation of this algorithm, suitable for log-linear models, in Section 3.2.

For incomplete contingency tables, let $\mathbf{y} = \left(\mathbf{y}^{(O)}, \mathbf{y}^{(U)}\right)$ where $\mathbf{y}^{(O)}$ and $\mathbf{y}^{(U)}$ are the observed and unobserved cell counts, respectively. Furthermore, let $\mathcal{O}$ and $\mathcal{U}$ denote the sets of observed and unobserved cells, respectively, so that $\mathcal{O} \cap \mathcal{U} = \emptyset$ and $\mathcal{O} \cup \mathcal{U} = \mathcal{S}$. We evaluate the joint posterior distribution of $\mathbf{y}^{(U)}$, ϕ, $\boldsymbol{\theta}_m$, σ^2, and m, i.e.,

$$\pi(\mathbf{y}^{(U)}, \phi, \boldsymbol{\theta}_m, \sigma^2, m|\mathbf{y}^{(O)}) \propto \pi(\mathbf{y}^{(O)}, \mathbf{y}^{(U)}|\phi, \boldsymbol{\theta}_m, m)\pi(\boldsymbol{\theta}_m|\sigma^2, m)\pi(\sigma^2)\pi(m), \tag{8}$$

where $\pi(\mathbf{y}^{(O)}, \mathbf{y}^{(U)}|\phi, \boldsymbol{\theta}_m, m) = \pi(\mathbf{y}|\phi, \boldsymbol{\theta}_m, m)$ is the complete-data likelihood function. We generate a sample from this posterior distribution using a data-augmentation MCMC algorithm (King and Brooks 2001). This algorithm is briefly described in Section 3.2.

In some cases, for incomplete contingency tables, one of the sources may observe individuals who are not members of the target population. An example of this is from the capture-recapture studies that have been used to estimate the number of people who inject drugs (PWID) in Scotland in the years 2003, 2006 and 2009 (see, King, Bird, Brooks, Hutchinson, and Hay 2005; King, Bird, Hutchinson, and Hay 2009; King, Bird, Overstall, Hay, and Hutchinson 2013, respectively). Here there are four sources who observe PWID. One of the sources is the Hepatitis C virus (HCV) database. This database does not actually observe PWID, but instead people who are newly diagnosed with the HCV and have injecting drug use as a historical risk factor. Assuming that these people are PWID results in over-estimation of the total population size of PWID in Scotland (see King *et al.* 2009 and Overstall *et al.* 2014). A modeling approach was adopted by Overstall *et al.* (2014) whereby people observed by the HCV database *and* another source are regarded as PWID, however the true number of PWID observed by *just* the HCV database is unknown but bounded from above by the count in this cell, i.e., the cell count is left censored. Let $\mathbf{y}^{(O)}$, $\mathbf{y}^{(U)}$ and $\mathbf{y}^{(C)}$ be the observed, unobserved and censored cell counts. Let $\mathbf{z}^{(C)}$ be the observed cell counts of the censored cells, i.e., the upper bound on the true cell counts, $\mathbf{y}^{(C)}$, for the censored cells. Furthermore, let $\mathcal{C}$ denote the set of all censored cells, so that all pairwise intersections of $\mathcal{O}$, $\mathcal{U}$ and $\mathcal{C}$ are the empty set and $\mathcal{O} \cup \mathcal{U} \cup \mathcal{C} = \mathcal{S}$. We evaluate the joint posterior distribution of $\mathbf{y}^{(U)}$, $\mathbf{y}^{(C)}$, ϕ, $\boldsymbol{\theta}_m$, σ^2 and m, given by

$$\pi(\mathbf{y}^{(U)}, \mathbf{y}^{(C)}, \phi, \boldsymbol{\theta}_m, \sigma^2, m|\mathbf{y}^{(O)}, \mathbf{z}^{(C)}) \propto \pi(\mathbf{y}^{(O)}, \mathbf{y}^{(U)}, \mathbf{y}^{(C)}|\phi, \boldsymbol{\theta}_m, m)\pi(\mathbf{z}^{(C)}|\mathbf{y}^{(C)})\pi(\boldsymbol{\theta}_m|\sigma^2, m)\pi(\sigma^2)\pi(m), \quad (9)$$

where $\pi(\mathbf{y}^{(O)}, \mathbf{y}^{(U)}, \mathbf{y}^{(C)}|\phi, \boldsymbol{\theta}_m, m) = \pi(\mathbf{y}|\phi, \boldsymbol{\theta}_m, m)$ is the complete-data likelihood function and $\pi(\mathbf{z}^{(C)}|\mathbf{y}^{(C)})$ gives the distribution of $\mathbf{z}^{(C)}$ conditional on $\mathbf{y}^{(C)}$. Overstall *et al.* (2014) specified

$$z_{\mathbf{i}}^{(C)}|y_{\mathbf{i}}^{(C)} \sim \mathrm{U}[y_{\mathbf{i}}^{(C)}, \infty),$$

for $\mathbf{i} \in \mathcal{C}$, i.e., uninformative censoring. The data-augmentation algorithm used to generate a sample from the posterior distribution given by (9) is described in Section 3.2.

The posterior distributions given by (7), (8) and (9) are all joint distributions of the model parameters, missing cell counts (in the case of (8) and (9) for incomplete contingency tables) and the model indicator. The model defined by such a joint distribution (i.e., including the model indicator) is known as the encompassing model (O'Hagan and Forster 2004, p. 164).

3.2. Computation

In this section we describe the computational algorithms used to generate an MCMC sample from the posterior distributions given by (7), (8) and (9) in Section 3.1, i.e., for complete contingency tables, and for incomplete contingency tables (without and with censoring). We begin by describing the most general data-augmentation algorithm which can be used to generate an MCMC sample from the posterior distribution given by (9), i.e., the joint posterior distribution of ϕ, $\boldsymbol{\theta}_m$, σ^2, m, $\mathbf{y}^{(U)}$ and $\mathbf{y}^{(C)}$. The algorithms for generating an MCMC sample from the posterior distributions given by (7) and (8) are special cases of the general algorithm.

The general data-augmentation algorithm cycles through the following steps.

1. Given the current values of ϕ, $\boldsymbol{\theta}_m$, σ^2, m and $\mathbf{y}^{(C)}$, generate new values of the unobserved cell counts, $\mathbf{y}^{(U)}$, from the full conditional distribution which is denoted by $\pi(\mathbf{y}^{(U)}|\phi, \boldsymbol{\theta}_m, \sigma^2, m, \mathbf{y}^{(C)}, \mathbf{y}^{(O)}, \mathbf{z}^{(C)})$.

2. Given the current values of ϕ, $\boldsymbol{\theta}_m$, σ^2, m and $\mathbf{y}^{(U)}$, generate new values of the censored cell counts, $\mathbf{y}^{(C)}$, from the full conditional distribution which is denoted by $\pi(\mathbf{y}^{(C)}|\phi, \boldsymbol{\theta}_m, \sigma^2, m, \mathbf{y}^{(U)}, \mathbf{y}^{(O)}, \mathbf{z}^{(C)})$.

3. Given the current values of $\mathbf{y}^{(C)}$ and $\mathbf{y}^{(U)}$, generate new values of the model parameters and model indicator, ϕ, $\boldsymbol{\theta}_m$, σ^2 and m, from the full conditional distribution which is denoted by $\pi(\phi, \boldsymbol{\theta}_m, \sigma^2, m|\mathbf{y}^{(C)}, \mathbf{y}^{(U)}, \mathbf{y}^{(O)}, \mathbf{z}^{(C)})$.

We will shortly describe how we can generate values from the full conditional distributions given in Steps 1, 2 and 3. The above algorithm is implemented in **conting** by the functions `bict()` and `bictu()`, which act as wrapper functions for `bict.fit()` which is the workhorse. For incomplete contingency tables, with no censored cell counts, we can skip Step 2. This algorithm is also implemented by `bict()` and `bictu()`. For complete contingency tables, we can skip Steps 1 and 2. This algorithm is implemented in **conting** by the functions `bcct()` and `bcctu()`, which again act as wrapper functions for `bcct.fit()`.

We now briefly describe how the simulation in each of the above three steps can be achieved. We do not describe the methods in detail but outline the algorithms involved and point the appropriate references out to the reader.

The simulation in Step 1 is trivial. The full conditional distribution can be written as

$$\pi(\mathbf{y}^{(U)}|\phi, \boldsymbol{\theta}_m, \sigma^2, m, \mathbf{y}^{(C)}, \mathbf{y}^{(O)}, \mathbf{z}^{(C)}) \propto \pi(\mathbf{y}^{(U)}|\phi, \boldsymbol{\theta}_m, m),$$

and each element of $\mathbf{y}^{(U)}$ can be generated as follows

$$y_{\mathbf{i}}|\phi, \boldsymbol{\theta}_m, m \sim \text{Poisson}\left(\mu_{\mathbf{i}}\right),$$

for $\log \mu_{\mathbf{i}} = \phi + \mathbf{x}_{m,\mathbf{i}}^{\top}\boldsymbol{\theta}_m$ and $\mathbf{i} \in \mathcal{U}$.

Similarly in Step 2, the full conditional distribution can be written as

$$\pi(\mathbf{y}^{(C)}|\phi, \boldsymbol{\theta}_m, \sigma^2, m, \mathbf{y}^{(U)}, \mathbf{y}^{(O)}, \mathbf{z}^{(C)}) \propto \pi(\mathbf{y}^{(C)}|\phi, \boldsymbol{\theta}_m, m)\pi(\mathbf{z}^{(C)}|\mathbf{y}^{(C)}),$$

and each element of $\mathbf{y}^{(C)}$ can be generated as follows

$$y_{\mathbf{i}}|\phi, \boldsymbol{\theta}_m, z_{\mathbf{i}}, m \sim \text{Trunc-Poisson}\left(\mu_{\mathbf{i}}, z_{\mathbf{i}}\right),$$

for $\mathbf{i} \in \mathcal{C}$, where $z_{\mathbf{i}}$ is the element of $\mathbf{z}^{(C)}$ corresponding to $y_{\mathbf{i}}$ and Trunc-Poisson(μ, z) is the truncated Poisson(μ) distribution bounded from above by z.

In Step 3, the univariate full conditional distribution of σ^2 is

$$\sigma^2|\boldsymbol{\theta}_m, m \sim \text{IG}\left(\frac{a + p_m}{2}, \frac{b + \frac{1}{n}\boldsymbol{\theta}_m^{\top}\mathbf{X}_m^{\top}\mathbf{X}_m\boldsymbol{\theta}_m}{2}\right).$$

The full conditional distribution of ϕ, $\boldsymbol{\theta}_m$ and m can be written as

$$\pi(\phi, \boldsymbol{\theta}_m, m|\mathbf{y}, \sigma^2) \propto \pi(\mathbf{y}|\phi, \boldsymbol{\theta}_m, m)\pi(\boldsymbol{\theta}_m|\sigma^2, m)\pi(m),$$

and can be generated from using the reversible jump algorithm. The reversible jump algorithm works as follows. Suppose the current model is m with current log-linear parameters $\boldsymbol{\beta}_m = (\phi, \boldsymbol{\theta}_m^\top)^\top$. We propose a move to model k with probability $\pi_{m,k}$. Innovation variables (Green 2003), which are denoted by $\mathbf{u}_m$, are generated from some proposal distribution and then a mapping function is applied to $\boldsymbol{\beta}_m$ and $\mathbf{u}_m$ to produce the proposed log-linear parameters $\boldsymbol{\beta}_k = (\phi, \boldsymbol{\theta}_k^\top)^\top$. We accept this proposed move with an associated acceptance probability. The key is the joint specification of the mapping function and the proposal distribution with different implementations resulting from different specifications. We use the specification for GLMs proposed by Forster, Gill, and Overstall (2012). For details on this specification for log-linear models, see Overstall *et al.* (2014). This method is implemented in **conting** by the function `RJ_update()`.

Note that $\pi_{m,m}$, the probability of proposing a move to the current model (called a null move), can be positive. In this case, we generate new values for the log-linear parameters, $\boldsymbol{\beta}_m$, conditional on the model, m. We use the Metropolis-Hastings algorithm, in particular, the iterated weighted least squares implementation for GLMs proposed by Gamerman (1997). For details on this method applied to log-linear models, see Overstall *et al.* (2014). This method is implemented in **conting** by the function `iwls_mh()`.

The values of $\pi_{m,k}$ are specified such that only local moves are proposed, i.e., $\pi_{m,k}$ are only non-zero for models k that can be derived from m by adding or dropping a single interaction term (both subject to the principle of marginality). The moves of adding or dropping a term are referred to as birth or death moves, respectively, by Forster *et al.* (2012). Suppose our current model is m, and we can propose a birth or death move to T_m models, then we set $\pi_{m,k} = (1 - \pi_{m,m})T_m^{-1}$ for $k \neq m$. In **conting**, the user may specify the probability, $\pi_{m,m}$, of the null move as an optional argument to the functions `bcct()` and `bict()`, with default value 0.5.

3.3. Assessing model adequacy

In this section we briefly review the method of assessing model adequacy, under the Bayesian approach, which is implemented in **conting**. This method uses predictions from the model of the observed cell counts. The idea is to compare these predicted cell counts to the observed cell counts. If they are inconsistent, then we can conclude that the model is inadequate. For more details on other approaches to assessing model adequacy, under the Bayesian approach, see Gelman *et al.* (2004, Chapter 6).

The comparison of predicted and observed cell counts can be made using the Bayesian (or posterior predictive) p value. In general, let $\mathbf{y}^{\text{rep}}$ and $\boldsymbol{\psi}$ denote the predicted cell counts and the vector of all of the model parameters. Define $T(\mathbf{Y}, \boldsymbol{\psi})$ to be a discrepancy statistic that can depend on both the cell counts, $\mathbf{Y}$, and the parameters, $\boldsymbol{\psi}$. The Bayesian p value is defined as

$$p_B = \mathsf{P}\left(T(\mathbf{y}^{\text{rep}}, \boldsymbol{\psi}) \geq T(\mathbf{y}, \boldsymbol{\psi}) | \mathbf{y}\right),$$

where the probability is with respect to the joint posterior distribution of $\mathbf{y}^{\text{rep}}$ and $\boldsymbol{\psi}$. If the model is inadequate, the distribution of $T(\mathbf{y}^{\text{rep}}, \boldsymbol{\psi})$ will be inconsistent with the distribution of $T(\mathbf{y}, \boldsymbol{\psi})$, therefore the Bayesian p value will be close to zero or one.

The Bayesian p value differs from the classical p value in that for a classical p value the discrepancy statistic only depends on the cell counts. The probability that defines the classical p value is with respect to $\mathbf{y}^{\text{rep}}$, conditional on a fixed value for $\boldsymbol{\psi}$ (either from the null

hypothesis or an estimated value). Under the Bayesian approach, the discrepancy statistic can depend on $\boldsymbol{\psi}$, over which we integrate out uncertainty with respect to the posterior distribution.

We can easily estimate the Bayesian p value using the MCMC sample generated from the posterior distribution. In the case of log-linear models applied to contingency tables, $\boldsymbol{\psi} = (\phi, \boldsymbol{\theta}_m, m)$, and let $\phi^{(j)}$, $\boldsymbol{\theta}^{(j)}_{m^{(j)}}$ and $m^{(j)}$ be the jth value of ϕ, $\boldsymbol{\theta}_m$ and m from the MCMC chain. We generate a predicted response vector, $\mathbf{y}^{(j)}$, using

$$y^{(j)}_{\mathbf{i}} | \phi^{(j)}, \boldsymbol{\theta}^{(j)}_{m^{(j)}}, m^{(j)} \sim \text{Poisson}\left(\mu^{(j)}_{\mathbf{i}}\right),$$

where $\log \mu^{(j)}_{\mathbf{i}} = \phi^{(j)} + \mathbf{x}^{\top}_{m^{(j)},\mathbf{i}} \boldsymbol{\theta}^{(j)}_{m^{(j)}}$, for $\mathbf{i} \in \mathcal{O}$. Therefore for an MCMC sample of size M, we will have M sampled values, $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ from the posterior predictive distribution of $\mathbf{y}^{\text{rep}}$. Let

$$\begin{aligned} T^{(j)}_O &= T(\mathbf{y}, \boldsymbol{\mu}^{(j)}), \\ T^{(j)}_P &= T(\mathbf{y}^{(j)}, \boldsymbol{\mu}^{(j)}), \end{aligned}$$

and we estimate the Bayesian p value to be the proportion of $T^{(j)}_P \geq T^{(j)}_O$, for $j = 1, \ldots, M$.

Examples of discrepancy statistics which are appropriate for contingency table data are the χ^2, deviance and Freeman-Tukey statistics, given by

$$\begin{array}{lcl} \chi^2 & : T(\mathbf{y}, \boldsymbol{\mu}) & = \sum_{\mathbf{i} \in \mathcal{O}} \left(y_{\mathbf{i}} - \mu_{\mathbf{i}}\right)^2 / \mu_{\mathbf{i}}, \\ \text{Deviance} & : T(\mathbf{y}, \boldsymbol{\mu}) & = -2 \sum_{\mathbf{i} \in \mathcal{O}} \left(y_{\mathbf{i}} \log \mu_{\mathbf{i}} - \mu_{\mathbf{i}} - \log y_{\mathbf{i}}!\right), \\ \text{FreemanTukey} & : T(\mathbf{y}, \boldsymbol{\mu}) & = \sum_{\mathbf{i} \in \mathcal{O}} \left(\sqrt{y_{\mathbf{i}}} - \sqrt{\mu_{\mathbf{i}}}\right)^2. \end{array}$$

Each of these statistics can be used within **conting** to assess model adequacy using the `bayespval()` function (see Section 3.4). This function estimates the Bayesian p value also and allows the user access to the sampled values $T^{(j)}_P$ and $T^{(j)}_O$, for $j = 1, \ldots, M$.

3.4. Functions of conting

Figure 1 shows all of the functions of **conting** that the user will typically call. The main two functions are `bcct()` and `bict()` that implement the MCMC algorithm in Section 3.2, for complete and incomplete contingency tables, respectively. Mandatory arguments to `bcct()` and `bict()` will be the form of the maximal model (in terms of a '`formula`' object) and the number of MCMC iterations to perform in the first instance. The data are introduced to `bcct()` and `bict()` using the `data` argument, where the data must be either a '`data.frame`' or '`table`' object. Otherwise, the variables in the maximal model are taken from the environment from which `bcct()` or `bict()` is called. Once the data-augmentation algorithm has finished the requested number of iterations, then additional iterations can be requested using `bcctu()` and `bictu()`, for `bcct()` and `bict()`, respectively.

The functions `bcct()` and `bcctu()` will return an object of class '`bcct`' which is a list containing all of the MCMC output. Correspondingly, `bict()` and `bictu()` will return an object of class '`bict`'. The S3 generics `print` and `summary` can be applied to '`bcct`' and '`bict`' objects producing a very simple summary (in the case of `print`) and a more detailed summary (in the case of `summary`).

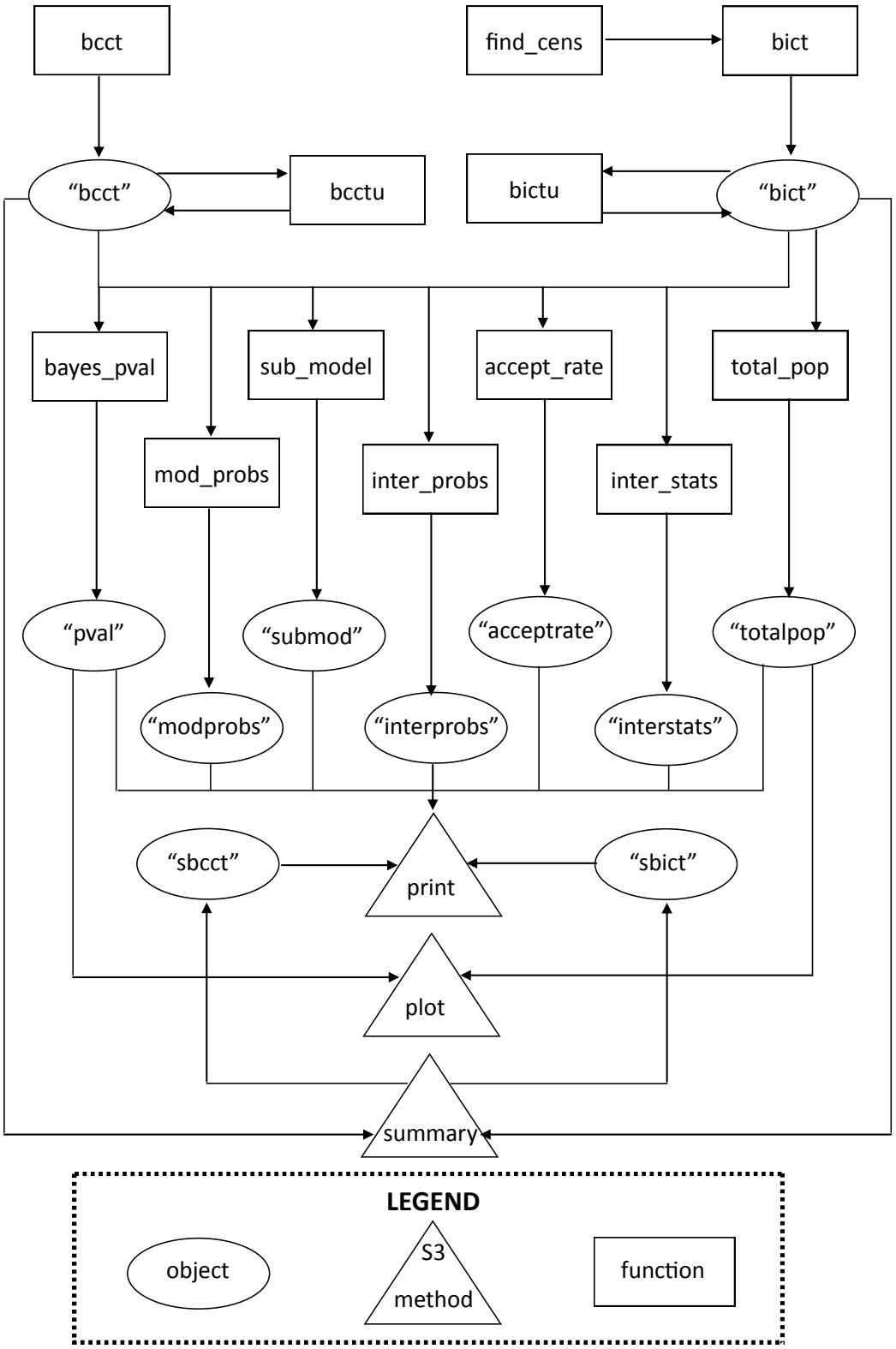

Figure 1: The functions of **conting**.

Name of function	Description	Produces an object of class	S3 methods that can be applied to this class
`accept_rate()`	Calculates the acceptance rates of the reversible jump and Metropolis-Hastings algorithms which can be used to assess MCMC performance (see Brooks, Giudici, and Roberts 2003).	'`acceptrate`'	`print`
`bayespval()`	Performs assessments of model adequacy by predicting from the model to calculate the Bayesian p value	'`pval`'	`print` `plot`
`inter_probs()`	Calculates the posterior probability of each term in the maximal model.	'`interprob`'	`print`
`inter_stats()`	Calculates the posterior probability, mean, variance, and highest posterior density interval (HPDI) of each log-linear parameter of the maximal model.	'`interstat`'	`print`
`mod_probs()`	Calculates the posterior model probabilities.	'`modprobs`'	`print`
`sub_model()`	Conditional on a user-specified model of interest, calculates the posterior mean, variance and HPDI for the log-linear parameters, as well as assessments of model adequacy and estimates of the total population size.	'`submod`'	`print`
`total_pop()`	Derives an MCMC sample from the posterior distribution of the total population size. Can only be applied to '`bict`' objects.	'`totpop`'	`print` `plot`

Table 1: The seven specific functions that can be used for '`bcct`' and '`bict`' objects.

There are also seven specific functions that can be used for '`bcct`' and '`bict`' objects and these are summarized in Table 1. The functions shown in Table 1 (except for `inter_probs()`, `accept_rate()` and `sub_model()`) are used to construct the output given by the `summary()` function applied to '`bcct`' and '`bict`' objects.

4. Examples

In this section we use four examples to demonstrate **conting**. The first two examples involve complete contingency tables and the latter two involve incomplete contingency tables. The data for the four examples are included in **conting**. In each example, we have specified the seed using the `set.seed()` function, so that all of the examples are fully reproducible.

4.1. Alcohol, obesity and hypertension

In Section 2, we introduced the AOH data as an example of a complete contingency table. We now conduct a Bayesian analysis of this table using the `bcct()` and `bcctu()` functions.

The mandatory arguments to `bcct()` are the form of the maximal model and the number of MCMC iterations. We specify that the maximal model be the saturated model which includes the three-way interaction. We initially request 1000 MCMC iterations. Optional arguments to `bcct()` involve specifying the starting values of the MCMC algorithm, the prior (Sabanés-Bové and Held, or unit information), the value of $\pi_{m,m}$ and details for saving the MCMC output to external files. Additionally, we can also specify the hyperparameters for the Sabanés-Bové and Held prior (a and b from Section 2) and request a progress bar to monitor the iterations. We assume the unit information prior distribution but allow the remaining arguments to take their default values so that the algorithm starts from the posterior mode of the maximal model, $\pi_{m,m} = 0.5$, the output is not saved to external files, and no progress bar is displayed. Once the initial 1000 MCMC iterations are complete we request a further 9000 iterations (making a total of 10000) using the `bcctu()` function. We then print out the resulting '`bcct`' object which gives a very simple summary which essentially informs the user of the analysis performed.

```
R> set.seed(1)
R> data("AOH", package = "conting")
R> aoh_ex <- bcct(formula = y ~ alc * hyp * obe, data = AOH,
+    n.sample = 1000, prior = "UIP")
R> aoh_ex <- bcctu(object = aoh_ex, n.sample = 9000)
R> aoh_ex

Number of cells in table = 24

Maximal model =
y ~ alc * hyp * obe

Number of log-linear parameters in maximal model = 24

Number of MCMC iterations = 10000

Computer time for MCMC = 00:01:06

Prior distribution for log-linear parameters = UIP
```

The object `aoh_ex` is a list which includes `BETA`, a matrix containing the MCMC samples for the log-linear parameters. We assess MCMC convergence informally (O'Hagan and Forster

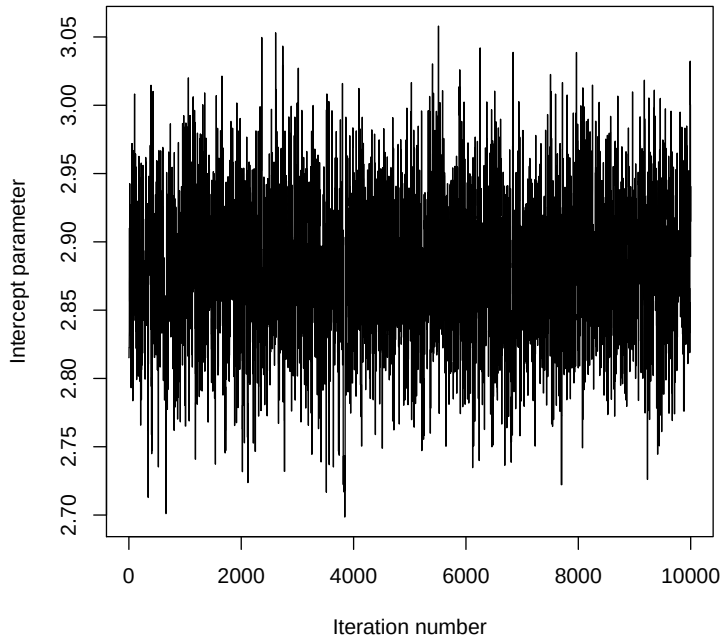

Figure 2: Trace plot for the intercept parameter, ϕ, in the alcohol, obesity and hypertension example.

2004, p. 426) by inspecting trace plots of the sampled values. For example we can produce a trace plot of the intercept parameter, ϕ, using the following line of code.

```
R> ts.plot(aoh_ex$BETA[, 1], ylab = "Intercept parameter",
+    xlab = "Iteration number")
```

The resultant plot is shown in Figure 2. We see that convergence occurs very quickly. For a detailed summary of the analysis performed, including posterior summary statistics, use the generic S3 method `summary()`. This function has optional arguments which control how the MCMC output is used. These are the number of burn-in iterations (`n.burnin`), the amount of thinning (`thin`), the cutoff of the posterior probability for presenting results on the log-linear parameters (`cutoff`), the discrepancy statistic (`statistic`), values to control which posterior model probabilities to present (`best` and `scale`), and the target probability (`prob.level`) for the highest posterior density intervals (HPDIs). We use a conservative burn-in phase of 5000 iterations, present posterior summary statistics only for log-linear parameters with probability greater than 0.05 and present the posterior model probabilities of the four models with the highest values for these probabilities. We use the χ^2 discrepancy statistic and no thinning which are the default values of the arguments `statistic` and `thin`.

```
R> aoh_ex_summ <- summary(aoh_ex, n.burnin = 5000, cutoff = 0.05, best = 4)
R> aoh_ex_summ

Posterior summary statistics of log-linear parameters:
            post_prob post_mean post_var lower_lim upper_lim
(Intercept)    1.0000   2.87831 0.002953   2.77321   2.98174
```

```
alc1           1.0000  -0.05246 0.007552  -0.22109   0.12462
alc2           1.0000  -0.06679 0.006965  -0.22473   0.10549
alc3           1.0000   0.09071 0.006147  -0.05021   0.24773
hyp1           1.0000  -0.51223 0.002758  -0.61998  -0.41760
obe1           1.0000  -0.03502 0.006740  -0.20669   0.11304
obe2           1.0000  -0.01829 0.004804  -0.15216   0.11661
hyp1:obe1      0.4556  -0.19692 0.005621  -0.33686  -0.05046
hyp1:obe2      0.4556  -0.03208 0.005590  -0.19581   0.09921
NB: lower_lim and upper_lim refer to the lower and upper values of the
95 % highest posterior density intervals, respectively

Posterior model probabilities:
  prob   model_formula
1 0.5344 ~alc + hyp + obe
2 0.4492 ~alc + hyp + obe + hyp:obe
3 0.0100 ~alc + hyp + obe + alc:hyp
4 0.0064 ~alc + hyp + obe + alc:hyp + hyp:obe

Total number of models visited =  4

Under the X2 statistic

Summary statistics for T_pred
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.183  18.810  23.250  23.950  28.080  61.940

Summary statistics for T_obs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.26   25.45   32.28   31.55   36.35   60.93

Bayesian p-value =  0.2246
```

First, under the χ^2 discrepancy statistic, the Bayesian p value of 0.22 does not indicate an inadequate model meaning that predictions of the cell counts are consistent with the observed cell counts. The posterior modal model is the independence model and the model with the second largest posterior model probability only contains the hypertension and obesity interaction (`hyp:obe`). These two models are estimated to account for approximately 98% of the posterior probability. This shows there are no strong interactions between the factors although there is some evidence for interaction between hypertension and obesity. From inspecting the posterior means of the log-linear parameters associated with the hypertension and obesity interaction, it indicates that as obesity level moves from low to high, the probability of hypertension increases.

Now suppose we were interested in inference solely based on the model with the second highest posterior model probability, i.e., the model with `formula = y ~ alc + hyp + obe + hyp:obe`. We can produce posterior summary statistics similar to those given by `summary()`, but which are conditional on a model of interest. This is accomplished using the `sub_model()` function. There are two ways to pass the model of interest to `sub_model()`; either by providing

a `formula`, or by the ranking of the model with respect to its posterior model probability (`order`). We demonstrate with the latter, i.e., set `order = 2`.

```
R> sub_model(object = aoh_ex, order = 2, n.burnin = 5000)

Posterior model probability =  0.4492

Posterior summary statistics of log-linear parameters:
            post_mean post_var lower_lim upper_lim
(Intercept)   2.86613 0.002993   2.76755   2.97230
alc1         -0.04840 0.007238  -0.21232   0.11952
alc2         -0.06900 0.006670  -0.24143   0.07961
alc3          0.09178 0.006074  -0.04605   0.24327
hyp1         -0.51954 0.002843  -0.61771  -0.41347
obe1         -0.08635 0.005781  -0.23917   0.05344
obe2         -0.02019 0.005841  -0.16255   0.13261
hyp1:obe1    -0.19665 0.005563  -0.33610  -0.05114
hyp1:obe2    -0.03199 0.005572  -0.19581   0.09921
NB: lower_lim and upper_lim refer to the lower and upper values of the
95 % highest posterior density intervals, respectively

Under the X2 statistic

Summary statistics for T_pred
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.662  18.730  23.330  23.950  28.350  63.310

Summary statistics for T_obs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.23   22.15   25.00   25.80   28.17   48.01

Bayesian p-value =  0.4114
```

Note that if the MCMC algorithm has not visited the specified model (given by `formula` or `order`) then `sub_model()` will return an informative error.

4.2. Risk factors for coronary heart disease

In this section, we consider the 2^6 complete contingency table given by Edwards and Havránek (1985) and used by, for example, Dellaportas and Forster (1999) and Forster *et al.* (2012) to demonstrate statistical methodology for complete contingency tables. Here 1841 men have been cross-classified by six risk factors (each with two levels) for coronary heart disease. The factors 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. We set the maximal model to be the saturated model including the six-way interaction. Starting from the model including all two-way interactions, we request 50000 iterations of the MCMC algorithm, under the unit information prior. We save the MCMC output to external files every 1000 iterations.

```
R> set.seed(1)
R> data("heart", package = "conting")
R> heart_ex <- bcct(formula = y ~ (A + B + C + D + E + F)^6, data = heart,
+    n.sample = 50000, start.formula = y ~ (A + B + C + D + E + F)^2,
+    save = 1000, prior = "UIP")
R> list.files()

[1] "BETA.txt"  "MHACC.txt" "MODEL.txt" "RJACC.txt" "SIG.txt"
```

The last line of code above, shows the MCMC output files in the working directory. We calculate the acceptance rates of the MCMC algorithms used.

```
R> accept_rate(heart_ex)

Acceptance rate of reversible jump proposals =  14.9384 %
Acceptance rate of Metropolis-Hastings proposals =  75.8558 %
```

Convergence of the MCMC iterations was assessed informally using trace plots (not shown) similar to Section 4.1. We now summarize the analysis. We use a burn-in phase of 10000 iterations, the deviance discrepancy statistic, and request 99% probability intervals for the log-linear parameters. We only present the posterior model probabilities for the models with the four highest probabilities. We could achieve this by using the `summary()` function. However, we instead calculate the Bayesian p value using the `bayespval()` function, the posterior summary statistics for the log-linear parameters using the `inter_stats()` function, and the posterior model probabilities using the `mod_probs()` function. Thus we demonstrate the use of these dedicated functions.

```
R> heart_ex_bp <- bayespval(heart_ex, n.burnin = 10000,
+    statistic = "deviance")
R> heart_ex_bp

Under the deviance statistic

Summary statistics for T_pred
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  312.0   344.7   352.0   352.6   359.9   416.1

Summary statistics for T_obs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  346.8   361.8   365.5   366.2   369.7   401.0

Bayesian p-value =  0.1502

R> inter_stats(heart_ex, n.burnin = 10000, prob.level = 0.99)

Posterior summary statistics of log-linear parameters:
```

```
            post_prob post_mean  post_var lower_lim upper_lim
(Intercept)    1.0000   2.69922 0.0015485   2.59372   2.79933
A1             1.0000  -0.03725 0.0007013  -0.10644   0.03646
B1             1.0000  -0.22973 0.0016583  -0.32235  -0.11395
C1             1.0000  -0.16069 0.0010445  -0.24495  -0.07997
D1             1.0000  -0.13602 0.0006591  -0.20335  -0.06653
E1             1.0000  -0.12742 0.0009243  -0.19963  -0.03890
F1             1.0000  -0.89965 0.0011254  -0.98633  -0.81541
A1:C1          1.0000   0.13400 0.0005913   0.07056   0.19688
A1:D1          0.9671  -0.08805 0.0005665  -0.14956  -0.02815
A1:E1          0.9992   0.11984 0.0006039   0.05499   0.18183
B1:C1          1.0000  -0.69960 0.0009270  -0.77933  -0.62331
D1:E1          0.9859   0.09479 0.0005992   0.03273   0.15784
NB: lower_lim and upper_lim refer to the lower and upper values of the
99 % highest posterior density intervals, respectively

R> mod_probs(heart_ex, n.burnin = 10000, best = 4)

Posterior model probabilities:
  prob
1 0.19462
2 0.12968
3 0.07010
4 0.06705
  model_formula
1 ~A + B + C + D + E + F + A:C + A:D + A:E + B:C + C:E + D:E
2 ~A + B + C + D + E + F + A:C + A:D + A:E + B:C + B:E + D:E
3 ~A + B + C + D + E + F + A:C + A:D + A:E + B:C + B:E + C:E + D:E
4 ~A + B + C + D + E + F + A:C + A:D + A:E + B:C + B:F + C:E + D:E

Total number of models visited =  394
```

Under the deviance discrepancy statistic, the Bayesian p value of 0.15 does not suggest an inadequate model. The models with the four highest posterior model probabilities are the same as identified by Dellaportas and Forster (1999) and Forster *et al.* (2012). With regards to the log-linear parameters we used the default value for `cutoff` so we are only presented with results for parameters with posterior probability greater than 0.75. This suggests there are strong interactions between smoking and strenuous physical work (`A:C`), systolic blood pressure (`A:D`) and the ratio of α and β lipoproteins (`A:E`), as well as for the interactions between strenuous mental and physical work (`B:C`) and between systolic blood pressure and the ratio of α and β lipoproteins (`D:E`).

4.3. Spina Bifida

In this section we consider a $2^3 \times 3$ incomplete contingency table given by Madigan and York (1997). Between 1969 and 1974, in upstate New York, 621 people born with Spina Bifida are observed by three sources: birth certificates (`S1`), death certificates (`S2`), and

medical rehabilitation records (S3). Additionally, these people are cross-classified according to ethnicity (eth) which has three levels: Caucasian, Afro-American and Other. These data can be used to estimate the total population size of people born with Spina Bifida by estimating the missing cell counts of people not observed by any of the three sources.

Below we print out the data. It can be seen that the three cell counts corresponding to people not observed by any of the three sources (one for each classification of ethnicity) are NA. This tells bict() that these cell counts are missing.

```
R> data("spina", package = "conting")
R> spina
```

```
     y  S1  S2  S3           eth
1   NA  un  un  un     caucasian
2   45  un obs  un     caucasian
3  230 obs  un  un     caucasian
4  134 obs obs  un     caucasian
5   52  un  un obs     caucasian
6    3  un obs obs     caucasian
7  107 obs  un obs     caucasian
8   12 obs obs obs     caucasian
9   NA  un  un  un afro-american
10   3  un obs  un afro-american
11  13 obs  un  un afro-american
12   8 obs obs  un afro-american
13   8  un  un obs afro-american
14   1  un obs obs afro-american
15   3 obs  un obs afro-american
16   0 obs obs obs afro-american
17  NA  un  un  un         other
18   0  un obs  un         other
19   1 obs  un  un         other
20   0 obs obs  un         other
21   0  un  un obs         other
22   0  un obs obs         other
23   1 obs  un obs         other
24   0 obs obs obs         other
```

We begin by considering the independence model only, i.e., we do not consider model uncertainty. Single model inference can be achieved using bict() (and bcct() for complete contingency tables) by specifying the model of interest as the maximal model and then setting $\pi_{m,m} = 1$ (i.e., null.move.prob = 1). Note that single model inference can also be achieved, having already fitted the encompassing model (Section 3), by using the sub_model() function as demonstrated in Section 4.1. We use 25000 iterations under the default Sabanés-Bové and Held prior. When the MCMC algorithm has finished, we summarize the analysis after discarding the first 5000 iterations as burn-in and use the default χ^2 discrepancy statistic.

```
R> set.seed(1)
R> spina_ex_ind <- bict(formula = y ~ S1 + S2 + S3 + eth, data = spina,
+    n.sample = 25000, null.move.prob = 1)
```

```
R> summary(spina_ex_ind, n.burnin = 5000)

Posterior summary statistics of log-linear parameters:
            post_prob post_mean post_var lower_lim upper_lim
(Intercept)         1   1.37569 0.053109    0.8693    1.7748
S11                 1  -0.35986 0.003265   -0.4731   -0.2501
S21                 1   0.49110 0.002028    0.4047    0.5798
S31                 1   0.55756 0.002120    0.4693    0.6495
eth1                1   2.77890 0.052561    2.3624    3.2639
eth2                1  -0.01643 0.057632   -0.4700    0.4627
NB: lower_lim and upper_lim refer to the lower and upper values of the
95 % highest posterior density intervals, respectively

Posterior model probabilities:
  prob model_formula
1 1    ~S1 + S2 + S3 + eth

Total number of models visited =  1

Posterior mean of total population size = 757.4334
95 % highest posterior density interval for total population size =
( 717 797 )

Under the X2 statistic

Summary statistics for T_pred
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.517  14.480  19.050  20.970  24.940 246.800

Summary statistics for T_obs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  61.21   68.07   71.92   73.22   76.75  110.60

Bayesian p-value =  0.003
```

The Bayesian p value is very close to zero which suggests that the predicted cell counts from the model are inconsistent with the observed cell counts indicating that the model is inadequate. We elaborate the model by including model uncertainty and fitting the encompassing model. We set the maximal model as the model with all two-way interactions and request 25000 iterations.

```
R> set.seed(1)
R> spina_ex <- bict(formula = y ~ (S1 + S2 + S3 + eth)^2, data = spina,
+    n.sample = 25000)
R> spina_ex_summ <- summary(spina_ex, n.burnin = 5000)
R> spina_ex_summ
```

```
Posterior summary statistics of log-linear parameters:
            post_prob post_mean post_var lower_lim upper_lim
(Intercept)         1   1.13984 0.065807    0.6276   1.57305
S11                 1  -0.42069 0.032310   -0.6791   0.05947
S21                 1   0.75397 0.008459    0.5912   0.91438
S31                 1   0.77893 0.007449    0.6151   0.94123
eth1                1   2.75560 0.062992    2.2954   3.21856
eth2                1   0.01322 0.077752   -0.4928   0.52171
S21:S31             1  -0.43865 0.004804   -0.5808  -0.30623
NB: lower_lim and upper_lim refer to the lower and upper values of the
95 % highest posterior density intervals, respectively

Posterior model probabilities:
  prob    model_formula
1 0.40965 ~S1 + S2 + S3 + eth + S2:S3
2 0.14595 ~S1 + S2 + S3 + eth + S1:S2 + S2:S3
3 0.13730 ~S1 + S2 + S3 + eth + S1:S3 + S2:S3
4 0.11350 ~S1 + S2 + S3 + eth + S1:eth + S2:S3
5 0.06110 ~S1 + S2 + S3 + eth + S1:S2 + S1:eth + S2:S3
6 0.05505 ~S1 + S2 + S3 + eth + S1:S3 + S1:eth + S2:S3

Total number of models visited =  24

Posterior mean of total population size = 725.8932
95 % highest posterior density interval for total population size =
( 676 781 )

Under the X2 statistic

Summary statistics for T_pred
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.509  13.970  18.240  20.930  24.190 745.100

Summary statistics for T_obs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.308  18.490  22.190  22.850  26.480  57.180

Bayesian p-value =  0.3622
```

The Bayesian p value is now 0.36 which indicates that there is no evidence of an inadequate model. By looking at the posterior model probabilities and the posterior probabilities of the log-linear parameters, there is overwhelming evidence of an interaction between the death certificate and medical rehabilitation sources. The posterior mean of the log-linear parameter associated with this interaction is negative indicating that a person observed by the death certificate source has a reduced probability of being observed by the medical rehabilitation source, and vice versa. The posterior mean of the total population size is 726. We can access the MCMC sample from the posterior distribution of the total population size using the

`total_pop()` function. The `total_pop()` function will return an object of class '`totalpop`' which is a list including a component called `TOT`, which is the required MCMC sample. Below, as an example, we use this sample to produce the point estimate that minimizes the relative squared error loss function, i.e., the point estimate is given by $\hat{N} = \mathsf{E}(N^{-1}|\mathbf{y}^{(O)})/\mathsf{E}(N^{-2}|\mathbf{y}^{(O)})$, (Madigan and York 1997).

```
R> spina_tot <- total_pop(spina_ex, n.burnin = 5000)
R> round(mean(1/spina_tot$TOT) / mean(1/(spina_tot$TOT^2)), 0)
```

```
[1] 724
```

Madigan and York (1997) found a point estimate of 731, under the same loss function, with a 95% probability interval of (689, 794). Note that they collapsed the Afro-American and Other levels of the ethnicity factor to one level: Afro-American and Other.

4.4. Estimating the number of people who inject drugs in Scotland

In this section we consider a 2^7 incomplete contingency table which is given by King *et al.* (2013). These data relate to estimating the number of PWID in Scotland in 2006. There are four sources: social inquiry reports (`S1`); hospital records (`S2`); Scottish Drug Misuse Database (`S3`); HCV database (`S4`), and three additional factors: age (`age`; $<$ 35 years, 35+ years), gender (`gender`), and region (`region`; Greater Glasgow & Clyde, Rest of Scotland). A total of 5670 PWID are observed by the four sources. As discussed in Section 3.1, the cell counts corresponding to only being observed by the HCV database (`S4`) are subject to censoring. We need to identify which cells are subject to censoring. This can be done via the `find_cens()` function.

```
R> data("ScotPWID", package = "conting")
R> cens <- find_cens(sources = ~ S1 + S2 + S3 + S4, cens_source = ~ S4,
+    data = ScotPWID)
R> ScotPWID[cens, ]
```

```
      y S1 S2 S3  S4 Region Gender   Age
9   122 un un un obs    GGC   Male Young
25  135 un un un obs    GGC   Male   Old
41   48 un un un obs    GGC Female Young
57   38 un un un obs    GGC Female   Old
73  134 un un un obs   Rest   Male Young
89  104 un un un obs   Rest   Male   Old
105  78 un un un obs   Rest Female Young
121  25 un un un obs   Rest Female   Old
```

We set the maximal model to include all two-way interactions and specify the default Sabanés-Bové and Held prior distribution. We request 2 million iterations of the MCMC algorithm and save the MCMC output to external files every 1000 iterations. We also tell the function `bict()` that we have some censored cells using the `cens` argument. The default value for this argument is `NULL` meaning there are no censored cell counts.

```
R> set.seed(1)
R> scot_ex <- bict(
+    formula = y ~ (S1 + S2 + S3 + S4 + Age + Gender + Region)^2,
+    data = ScotPWID, cens = cens, n.sample = 2000000, save = 1000)
```

Throughout our summary of the above analysis we use a burn-in phase of 200000 iterations (i.e., we discard the first 10% of the iterations) and due to the size of the MCMC output we base our inference only on every 5th iteration by setting `thin = 5`. Following Gelman *et al.* (2004, p. 295) we only recommend using thinning when there are problems with computer memory due to large numbers of parameters and/or MCMC iterations. To assess model adequacy we use the Freeman-Tukey discrepancy statistic.

```
R> scot_ex_bp <- bayespval(scot_ex, n.burnin = 200000, thin = 5,
+    statistic = "FreemanTukey")
R> scot_ex_bp

Under the Freeman-Tukey statistic

Summary statistics for T_pred
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.23   28.49   31.30   31.54   34.32   54.88

Summary statistics for T_obs
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.13   33.00   34.24   34.44   35.67   50.62

Bayesian p-value =  0.26
```

The Bayesian p value is 0.26 which indicates that there is no evidence of an inadequate model. We now calculate the posterior probabilities of the log-linear parameters using the `inter_probs()` function and only present those with probability greater than 0.70.

```
R> inter_probs(scot_ex, n.burnin = 200000, thin = 5, cutoff = 0.70)

Posterior probabilities of log-linear parameters:
            post_prob
(Intercept)    1.0000
S1             1.0000
S2             1.0000
S3             1.0000
S4             1.0000
Age            1.0000
Gender         1.0000
Region         1.0000
S1:S3          0.9442
S1:Age         0.9991
S2:S4          1.0000
```

```
S2:Age          0.9865
S2:Gender       0.9999
S3:Age          1.0000
S3:Region       1.0000
Age:Gender      1.0000
Age:Region      1.0000
```

The posterior probabilities of the log-linear parameters are consistent with those found by Overstall *et al.* (2014). Finally we derive an MCMC sample from the posterior distribution of the total population size and find its posterior mean and 95% HPDI using the `tot_pop()` function.

```
R> scot_ex_tot <- total_pop(scot_ex, n.burnin = 200000, thin = 5)
R> scot_ex_tot

Posterior mean of total population size = 22856.24
95 % highest posterior density interval for total population size =
( 16427 27097 )
```

The posterior mean of the total population size of PWID in Scotland in 2006 is 22900 (to nearest 100). This value is consistent with that found by Overstall *et al.* (2014). Applying the S3 generic `plot()` to an object of class '`totpop`' will produce a histogram of the MCMC sample from the posterior distribution of the total population size.

```
R> plot(scot_ex_tot)
```

The resultant histogram is shown in Figure 3. Note that the bimodal nature of the posterior distribution of the total population size was also found by Overstall *et al.* (2014). This bimodality is caused by the presence, or absence, of the interaction between the social inquiry and Scottish Drug Misuse Database sources (`S1:S3`). The posterior mean for this interaction is positive so that the upper mode corresponds to the presence of the interaction, and the lower mode to absence.

5. Concluding remarks

This paper demonstrates the use of the R package **conting**, for the Bayesian analysis of complete and incomplete contingency tables. The **conting** package allows a user to identify interactions between factors and to estimate closed populations using incomplete contingency tables from capture-recapture studies. The capabilities of **conting** are demonstrated via four examples.

A novice user need only supply the data, the form of the maximal model (in the usual R way of using a `formula`), and the number of MCMC iterations. The S3 method for `summary()` will then provide all of the relevant information.

However a more experienced user can also access the sampled values of the model parameters and missing cell counts (for an incomplete table). For example, from Section 4.4, `scot_ex$BETA` is the 2 million by 29 matrix containing the sampled values of the $\boldsymbol{\beta}$ parameters and `scot_ex$Y0` is the 2 million by 16 matrix containing the sampled values of the

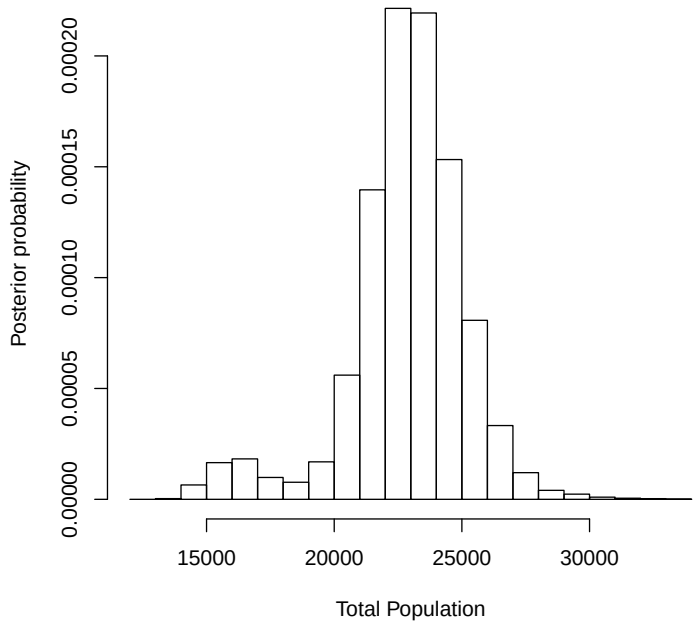

Figure 3: Histogram of the posterior distribution of the total population size of PWID in Scotland in 2006.

unobserved and censored cell counts. Furthermore, if the argument `save` is not `NULL`, then `bcct()` and `bict()` will save the MCMC output to external files; see Section 4.2. These files can be read into R and manipulated however the user wishes. Possible applications of this output include more extensive assessments of MCMC convergence and model adequacy, and the calculation of posterior summaries relevant to a specific research question.

Future work will involve updating **conting** to fit Bayesian models to contingency tables involving structural zeros, misclassified cell counts and ordinal factor levels.

Acknowledgments

The authors thank the reviewers for their constructive comments and suggestions, which have greatly improved the article and package. Both authors were partly funded by the MRC-funded addictions cluster, NIQUAD (Grant No. G1000021).

References

Agresti A (2007). *An Introduction to Categorical Data Analysis*. 2nd edition. John Wiley & Sons.

Brooks SP, Giudici P, Roberts GO (2003). "Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions." *Journal of the Royal Statistical Society B*, **65**(1), 3–55.

Dawid AP, Lauritzen SL (1993). "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models for Contingency Tables." *The Annals of Statistics*, **21**(3), 1272–1317.

Dellaportas P, Forster JJ (1999). "Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-Linear Models." *Biometrika*, **86**(3), 615–633.

Edwards D, Havránek T (1985). "A Fast Procedure for Model Search in Multidimensional Contingency Tables." *Biometrika*, **72**(2), 339–351.

Fienberg SE (1972). "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika*, **59**(3), 591–603.

Forster JJ (2010). "Bayesian Inference for Poisson and Multinomial Log-Linear Models." *Statistical Methodology*, **7**(3), 210–224.

Forster JJ, Gill RC, Overstall AM (2012). "Reversible Jump Methods for Generalised Linear Models and Generalised Linear Mixed Models." *Statistics and Computing*, **22**(1), 107–120.

Fox J (2002). *An R and S-PLUS Companion to Applied Regression*. Sage.

Gamerman D (1997). "Sampling from the Posterior Distribution in Generalised Linear Mixed Models." *Statistics and Computing*, **7**(1), 57–68.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis*. Chapman and Hall.

Green PJ (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, **82**(4), 711–732.

Green PJ (2003). "Trans-Dimensional Markov Chain Monte Carlo." In PJ Green, NL Hjort, S Richardson (eds.), *Highly Structured Stochastic Systems*. Oxford University Press.

Kass RE, Wasserman L (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, **91**(435), 1343–1370.

King R, Bird SM, Brooks SP, Hutchinson SJ, Hay G (2005). "Prior Information in Behavioural Capture-Recapture Methods: Demography Influences Injectors' Propensity to Be Listed on Data-Sources and their Drugs-Related Mortality." *American Journal of Epidemiology*, **162**(7), 694–703.

King R, Bird SM, Hutchinson SJ, Hay G (2009). "Estimating Current Injectors in Scotland and Their Drug-Related Death Rate by Sex, Region and Age-Group via Bayesian Capture-Recapture Methods." *Statistical Methods in Medical Research*, **18**(4), 341–359.

King R, Bird SM, Overstall AM, Hay G, Hutchinson SJ (2013). "Injecting Drug Users in Scotland, 2006: Number, Demography, and Opiate-Related Death-Rates." *Addiction Research and Theory*, **21**(3), 235–246.

King R, Brooks SP (2001). "On the Bayesian Analysis of Population Size." *Biometrika*, **88**(2), 317–336.

Knuiman MW, Speed TP (1988). "Incorporating Prior Information into the Analysis of Contingency Tables." *Biometrics*, **44**(4), 1061–1071.

Madigan D, York Y (1997). "Bayesian Methods for Estimation of the Size of a Closed Population." *Biometrika*, **84**(1), 19–31.

Mantel N (1970). "Incomplete Contingency Tables." *Biometrics*, **26**(2), 291–304.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall.

Ntzoufras I, Dellaportas P, Forster JJ (2003). "Bayesian Variable and Link Determination for Generalised Linear Models." *Journal of Statistical Planning and Inference*, **111**(1–2), 165–180.

O'Hagan A, Forster JJ (2004). *Kendall's Advanced Theory of Statistics*, volume 2B: Bayesian Inference. 2nd edition. John Wiley & Sons.

Overstall AM (2014). ***conting**: Bayesian Analysis of Contingency Tables*. R package version 1.3, URL http://CRAN.R-project.org/package=conting.

Overstall AM, King R (2014). "A Default Prior Distribution for Contingency Tables with Dependent Factor Levels." *Statistical Methodology*, **16**, 90–99.

Overstall AM, King R, Bird SM, Hutchinson SJ, Hay G (2014). "Incomplete Contingency Tables with Censored Cells with Application to Estimating the Number of People who Inject Drugs in Scotland." *Statistics in Medicine*, **33**(9), 1564–1579.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sabanés-Bové D, Held L (2011). "Hyper-g Priors for Generalized Linear Models." *Bayesian Analysis*, **6**(3), 387–410.

Tan M, Tian GL, Ng KW (2010). *Bayesian Missing Data Problems*. Chapman & Hall/CRC.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. URL http://www.stats.ox.ac.uk/pub/MASS4/.

Zellner A (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions." In PK Goel, A Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. North-Holland/Elsevier.

Affiliation:

Antony M. Overstall
School of Mathematics and Statistics
University of St Andrews

St Andrews
United Kingdom
E-mail: antony@mcs.st-and.ac.uk

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 58, Issue 7 *Submitted:* 2012-12-12
June 2014 *Accepted:* 2014-02-12