Statistical Methodology 16 (2014) 90–99

Contents lists available at ScienceDirect

Statistical Methodology

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

A default prior distribution for contingency tables with dependent factor levels[☆]

Antony M. Overstall [*], Ruth King

School of Mathematics & Statistics, University of St Andrews, St Andrews, Fife, KY16 9SS, United Kingdom

ARTICLE INFO

Article history:
Received 19 April 2013
Received in revised form
8 August 2013
Accepted 14 August 2013

Keywords:
Contingency table
Dependence structure
Default prior

ABSTRACT

A default prior distribution is proposed for the Bayesian analysis of contingency tables. The prior is specified to allow for dependence between levels of the factors. Different dependence structures are considered, including conditional autoregressive and distance correlation structures. To demonstrate the prior distribution, a dataset is considered which involves estimating the number of injecting drug users in the eleven National Health Service board regions of Scotland using an incomplete contingency table where the dependence structure relates to geographical regions.

1. Introduction

Contingency tables (e.g. [1]) are formed when a population is cross-classified according to a series of categories (or factors). Each cell count of the table gives the number observed under each cross-classification. The aim of forming such a table is to summarise the data, and typically, with a view to identifying interactions or relationships between the factors.

The standard statistical practice to model such interactions is the log–linear model (e.g. [1, Chapter 7]). In this case the logarithm of the expected cell count is proportional to a linear predictor depending on the main effect terms and interaction terms between the factors. Each combination of interaction terms defines its own log–linear model so that the identification of the non-zero interaction terms translates to an exercise in model comparison. Additionally incomplete contingency tables with missing cell counts can be used to estimate closed populations [4] where some of the factors correspond to sources that have either observed or not observed individuals in the population.

[*] Corresponding author. Tel.: +44 1344461806.
E-mail address: antony@mcs.st-and.ac.uk (A.M. Overstall).

1572-3127/$ – see front matter © 2014 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.stamet.2013.08.007

In this paper, we consider the case where the levels of one or more of the factors may be dependent on one another. An obvious example is when one of the factors has levels corresponding to geographical regions or locations which may be dependent due to their geographical proximity. In these cases, we may expect the parameters of the log–linear model to have some dependence structure. Bayesian analysis of contingency tables is common (e.g. [3,13,5]) and is the approach taken here. One feature of the Bayesian approach is that prior information on the interaction terms can be incorporated through the prior distribution. We take the position of having weak prior information on the magnitude of the log–linear parameters but wish to incorporate the information provided by the dependence structure mentioned above. In the case of weak prior information and model uncertainty, care must be taken when specifying prior distributions due to Lindley's paradox (e.g. [16, pp. 77–79]). There have been several attempts in the literature (e.g. [3,15,18]) to specify "default" prior distributions that can be applied for log–linear models under model uncertainty. We extend these approaches by developing a default prior that can take account of the dependence structure between the factor levels and can be seen as a generalisation of the above mentioned priors. The proposed prior is constructed by conditioning on the constraints on the parameters which are introduced in contingency table analysis to maintain identifiability of the parameters.

This paper is organised as follows. In Section 2 we set out our notation and briefly describe log–linear models. In Section 3 we derive our proposed default prior distribution including descriptions of different dependence structures. Finally, we apply our proposed prior to a real data application in Section 4, which involves estimating the number of injecting drug users in Scotland. Here, one of the factors corresponds to geographical regions, and we wish to take account of the possible dependence structure that may exist for the regions.

2. Notation and log–linear models

2.1. Notation

We assume that there are a total of c factors such that each factor $k = 1, \ldots, c$ has l_k levels. The corresponding contingency table has $n = \prod_{k=1}^{c} l_k$ cells. Let $\mathbf{y}$ be the $n \times 1$ vector of cell counts with elements denoted as $y_{\mathbf{i}}$ and where $\mathbf{i} = (i_1, \ldots, i_c)$ identifies the combination of factor levels that cross-classify the cell $\mathbf{i}$. Let $\mathcal{S}$ be set of all n cross-classifications so that

$$\mathcal{S} = \{(i_1, \ldots, i_c) : i_l \in \{1, \ldots, l_k\}\}.$$

Finally, let $N = \sum_{\mathbf{i} \in \mathcal{S}} y_{\mathbf{i}}$ be the total population size. In the case of an incomplete contingency table, N is unknown, since elements of $\mathbf{y}$ are unknown.

As a pedagogic example that we use for illustrative purposes throughout, suppose that there are three factors used to cross-classify a population of hospital patients: age (2 levels: young; old), hypertension (2 levels: no; yes) and region (3 levels: A; B; C). In this example, $c = 3$, where $l_1 = 2$, $l_2 = 2$ and $l_3 = 3$, and the three factors (age, hypertension and region) have been labelled 1, 2 and 3, respectively. It follows that there are $n = 2 \times 2 \times 3 = 12$ cells.

2.2. Log–linear models

We now briefly describe log–linear models and initially assume that the form of the log–linear model is known, i.e. it is known which interactions are present. We extend to the case of model uncertainty later in this section. Let $\eta_{\mathbf{i}}$ denote the linear predictor associated with cell $\mathbf{i} \in \mathcal{S}$, where

$$\eta_{\mathbf{i}} = \phi + \mathbf{z}_{\mathbf{i}}^T \boldsymbol{\theta},$$

with $\phi \in \mathbb{R}$ denoting the intercept term, $\boldsymbol{\theta}$ the $q \times 1$ vector of log–linear parameters (i.e. the main effects and interaction terms) and $\mathbf{z}_{\mathbf{i}}$ the $q \times 1$ vector of zeros and ones identifying which elements of $\boldsymbol{\theta}$ are applicable to cell $\mathbf{i} \in \mathcal{S}$.

For identifiability, certain elements of $\boldsymbol{\theta}$ are constrained, e.g. by sum-to-zero, or corner-point constraints, so we can rewrite $\eta_{\mathbf{i}}$ as

$$\eta_{\mathbf{i}} = \phi + \mathbf{x}_{\mathbf{i}}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the $p \times 1$ vector of unconstrained regression parameters, and $\mathbf{x_i}$ is the $p \times 1$ vector which identifies which elements of $\boldsymbol{\beta}$ correspond to cell $\mathbf{i} \in \mathcal{S}$, with $p < q$.

Finally, let $\boldsymbol{\eta}$ be the $n \times 1$ vector with elements $\eta_{\mathbf{i}}$, and let $\mathbf{X}$ be the $n \times p$ model matrix with rows $\mathbf{x}_i$. Then we can write

$$\boldsymbol{\eta} = \phi \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{1}_n$ denotes the $n \times 1$ vector of ones.

For the statistical analysis of contingency tables, it is common to assume that

$$y_{\mathbf{i}} | \phi, \boldsymbol{\beta} \sim \text{Poisson}\left(\lambda_{\mathbf{i}}\right), \tag{1}$$

independently, where $\log \lambda_{\mathbf{i}} = \eta_{\mathbf{i}}$.

In practice, we typically do not know the form of the log–linear model. This is equivalent to not knowing the elements of $\mathbf{z_i}$ and $\mathbf{x_i}$, or the columns of $\mathbf{X}$. Let $\mathcal{M}$ be the set of competing log–linear models which are indexed by $m \in \mathcal{M}$. Associated with each log–linear model are $\mathbf{z}_{\mathbf{i}}^{(m)}$, $\mathbf{x}_{\mathbf{i}}^{(m)}$, $\mathbf{X}^{(m)}$, $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\beta}^{(m)}$, where $\mathbf{z}_{\mathbf{i}}^{(m)}$ and $\boldsymbol{\theta}^{(m)}$ are $q^{(m)} \times 1$ vectors, $\mathbf{x}_{\mathbf{i}}^{(m)}$ and $\boldsymbol{\beta}^{(m)}$ are $p^{(m)} \times 1$ vectors, and $\mathbf{X}^{(m)}$ is an $n \times p^{(m)}$ matrix.

In the next section, we derive a default prior distribution for $\boldsymbol{\beta}^{(m)}|m$. For the intercept, ϕ, we assume a prior given by $\pi(\phi) \propto 1$. Although this prior is improper, the resulting posterior is still proper [5]. This prior will not cause a problem under Lindley's paradox since it is present for all models in $\mathcal{M}$ [16, p. 174].

3. A default prior distribution for $\boldsymbol{\beta}^{(m)}|m$

3.1. Derivation

In this section we develop a default prior distribution for $\boldsymbol{\beta}^{(m)}|m$. For notational simplicity, we drop the dependency on the model m by removing the superscript (m).

Suppose that there are a total of T log–linear terms and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\right)$ where $\boldsymbol{\beta}_t$, for $t = 1, \ldots, T$, is the $p_t \times 1$ vector corresponding to the regression parameters for the main effect or interaction term t. Similarly let $\boldsymbol{\theta}_t$ denote the corresponding $q_t \times 1$ vector of log–linear parameters, for $t = 1, \ldots, T$.

Let $\mathcal{R}_t$ be the set of f main effect terms that define the f-way interaction $\boldsymbol{\beta}_t$. Dellaportas and Forster [3] refer to $\mathcal{R}_t$ as the constituent terms of the interaction. Note that $q_t = \prod_{j \in \mathcal{R}_t} q_j$ and if $\boldsymbol{\beta}_t$ corresponds to a main effect then $\mathcal{R}_t$ has only one element, i.e. t. Consider the pedagogic example, from Section 2.1, and t corresponding to the 2-way interaction between age and region so that $q_t = 6$ and $p_t = 2$. The constituent terms, $\mathcal{R}_t$, have two elements: the terms corresponding to age and region.

We initially consider deriving the default prior distribution under sum-to-zero constraints. We describe how the prior can be extended to any system of constraints in Section 3.4. Following Dellaportas and Forster [3] we assume that $\boldsymbol{\beta}$ has a multivariate normal distribution with mean zero, where $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}_t$ are independent for $r, t = 1, \ldots, T$ and $r \neq t$. Thus, all that remains is to specify the $p_t \times p_t$ covariance matrix for each $\boldsymbol{\beta}_t$, for $t = 1, \ldots, T$.

The elements of $\boldsymbol{\theta}_t$ are subject to constraints and can be written in the form

$$\boldsymbol{\theta}_t = \mathbf{A}_t \boldsymbol{\beta}_t, \tag{2}$$

where $\mathbf{A}_t$ is a $q_t \times p_t$ matrix defining the constraints. Under sum-to-zero constraints, $\mathbf{A}_t$ can be written as

$$\mathbf{A}_t = \mathbf{P}_t \begin{pmatrix} \mathbf{I}_{p_t} \\ \mathbf{C}_t \end{pmatrix}, \tag{3}$$

where $\mathbf{I}_{p_t}$ is the $p_t \times p_t$ identity matrix, $\mathbf{C}_t$ is a $(q_t - p_t) \times p_t$ matrix and $\mathbf{P}_t$ is a $q_t \times q_t$ permutation matrix. For t corresponding to the age and region interaction in the pedagogic example,

$$\boldsymbol{\theta}_t = \begin{pmatrix} \theta_{t1} \\ \theta_{t2} \\ \theta_{t3} \\ \theta_{t4} \\ \theta_{t5} \\ \theta_{t6} \end{pmatrix}, \quad \mathbf{A}_t = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{C}_t = \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix},$$

$$\mathbf{P}_t = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

The elements of $\boldsymbol{\theta}_t$ are ordered so that the factor levels of region vary the fastest.

Initially, ignoring the constraints that are applied to $\boldsymbol{\theta}_t$, we assume that the distribution of $\boldsymbol{\theta}_t$ is

$$\boldsymbol{\theta}_t | \sigma_t^2, \mathbf{D}_t \sim \mathrm{N}\left(\mathbf{0}, \sigma_t^2 \mathbf{D}_t\right),$$

where $\sigma_t^2 > 0$ and $\mathbf{D}_t$ is a $q_t \times q_t$ positive-definite scale matrix. The off-diagonal elements of $\mathbf{D}_t$ control the dependence structure or correlation between the elements of the constrained parameters, $\boldsymbol{\theta}_t$, corresponding to different factor levels.

It follows from (2) and (3) that

$$\mathbf{P}_t^T \boldsymbol{\theta}_t = \begin{pmatrix} \boldsymbol{\beta}_t \\ \mathbf{C}_t \boldsymbol{\beta}_t \end{pmatrix}. \tag{4}$$

Let

$$\boldsymbol{\gamma}_t = \begin{pmatrix} \boldsymbol{\gamma}_t^{(1)} \\ \boldsymbol{\gamma}_t^{(2)} \end{pmatrix} = \mathbf{P}_t^T \boldsymbol{\theta}_t$$

be the permuted elements of $\boldsymbol{\theta}_t$ according to the inverse permutation $\mathbf{P}_t^{-1} = \mathbf{P}_t^T$, so that $\boldsymbol{\gamma}^{(1)} = \boldsymbol{\beta}_t$ and $\boldsymbol{\gamma}^{(2)} = \mathbf{C}_t \boldsymbol{\beta}_t$. The prior distribution for $\boldsymbol{\beta}_t$ is the conditional distribution of $\boldsymbol{\gamma}^{(1)}$ (which is $\boldsymbol{\beta}_t$) given that $\boldsymbol{\gamma}^{(2)} = \mathbf{C}_t \boldsymbol{\beta}_t$, i.e. we find the distribution of $\boldsymbol{\beta}_t$ from (4) subject to the constraints. It can be shown (see Appendix A) that

$$\boldsymbol{\beta}_t | \sigma_t^2, \mathbf{D}_t \sim \mathrm{N}\left(\mathbf{0}, \sigma_t^2 \boldsymbol{\Sigma}_t\right), \tag{5}$$

where

$$\boldsymbol{\Sigma}_t = \left(\mathbf{A}_t^T \mathbf{D}_t^{-1} \mathbf{A}_t\right)^{-1}. \tag{6}$$

In the next two sections we consider $\mathbf{D}_t$. It may be that $\mathbf{D}_t$ is completely specified *a priori*. The most plausible situation for this is when we assume independence between the levels of this term and $\mathbf{D}_t = \mathbf{I}_{q_t}$. We consider this case in Section 3.2. In Section 3.3 we also consider where $\mathbf{D}_t$ is unknown due to its dependence on some unknown hyperparameter which controls the strength of correlation between the elements of $\boldsymbol{\theta}_t$.

3.2. Independent correlation structure

Suppose we assume that the factor levels are independent, i.e. $\mathbf{D}_t = \mathbf{I}_{q_t}$, so that

$$\boldsymbol{\Sigma}_t = \left(\mathbf{A}_t^T \mathbf{A}_t\right)^{-1}.$$

Denote by $\mathbf{X}_t$ the $n \times p_t$ matrix formed by the columns of $\mathbf{X}$ corresponding to $\boldsymbol{\beta}_t$. Since $\mathbf{X}_t$ is a permutation of the matrix formed by stacking $\mathbf{A}_t$ to form an $n \times p_t$ matrix, it follows that

$$\mathbf{X}_t^T\mathbf{X}_t = \frac{n}{q_t}\mathbf{A}_t^T\mathbf{A}_t,$$

and therefore $\boldsymbol{\Sigma}_t = (n/q_t)\left(\mathbf{X}_t^T\mathbf{X}\right)^{-1}$. The corresponding prior distribution for $\boldsymbol{\beta}_t$ is

$$\boldsymbol{\beta}_t|\sigma_t^2 \sim \mathrm{N}\left(0, \frac{\sigma_t^2 n}{q_t}\left(\mathbf{X}_t^T\mathbf{X}_t\right)^{-1}\right).$$

If $\sigma_t^2 = gq_t/n$, then since (under sum-to-zero constraints) $\mathbf{X}_t^T\mathbf{X}_r \neq 0$, for all $t \neq r$ [14], it follows that the prior distribution for $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T\right)$ is

$$\boldsymbol{\beta}|g \sim \mathrm{N}\left(0, g\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right). \tag{7}$$

If $g > 0$ is unknown and given a prior distribution, then (7) is a hierarchical prior distribution that is identical to the generalised hyper-g prior proposed by Sabanes-Bové and Held [18] for generalised linear models (GLMs) when applied to log–linear models. If, instead, g is fixed then (7) is the default prior distribution considered by Dellaportas and Forster [3] who advocate setting $g = kn$ for some constant k, which represents the number of units of prior information. Ntzoufras et al. [15] use $k = 1$ under their unit information prior for GLMs when applied to log–linear models.

3.3. General correlation structure

We now consider terms, t, whose constituent terms, $\mathcal{R}_t$, contain factors with correlated levels and $\mathbf{D}_t$ depends on some unknown hyperparameter τ. This hyperparameter, τ, controls the strength of correlation through some structure imposed on $\mathbf{D}_t$. Initially consider a main effect term t. In this paper we focus on the case where the factor levels correspond to geographical regions or locations and propose two structural forms for $\mathbf{D}_t$. However there exist many possible applications with correlated factor levels and other correlation structures that can be used depending on the nature of the factor levels.

1. *Conditional autoregressive structure*
   Suppose that the q_t levels correspond to regions. Let $\mathbf{G}$ be the $q_t \times q_t$ neighbourhood matrix with ijth element

   $$G_{ij} = \begin{cases} 1 & \text{if regions } i \neq j \text{ are neighbours,} \\ 0 & \text{if otherwise,} \end{cases}$$

   for $i, j = 1, \ldots, q_t$. Then for the conditional autoregressive (CAR) structure (e.g. [2]),

   $$\mathbf{D}_t = \left(\mathbf{I}_{q_t} - \tau\mathbf{G}\right)^{-1},$$

   where τ determines the strength of spatial correlation for the constrained parameters. To ensure that $\mathbf{D}_t$ is positive-definite, the hyperparameter τ must lie in the interval $(\tau_{\min}, \tau_{\max}) = \left(e_{q_t}^{-1}, e_1^{-1}\right)$, where e_1 and e_{q_t} are the maximum and minimum eigenvalues of $\mathbf{G}$, respectively.
2. *Distance correlation structure*
   Suppose the q_t levels correspond to locations such as cities. Then the ijth element of $\mathbf{D}_t$ is given by a correlation function that depends on the distance, d_{ij}, between locations i and j, and τ. For example, the Gaussian correlation function gives

   $$D_{t,ij} = \exp\left(-\frac{d_{ij}^2}{2\tau^2}\right),$$

   where, again, $\tau > 0$ controls the strength of correlation.

Note that in both examples, the hyperparameter, τ, is not actually a correlation coefficient; it merely controls the strength of correlation. We need to specify a prior distribution for τ. This will depend on the application.

For a term t that corresponds to an interaction term, we propose

$$\mathbf{D}_t = \bigotimes_{r \in \mathcal{R}_t} \mathbf{D}_r. \tag{8}$$

The form given by (8) has been chosen for its consistency. Suppose that the correlation between two levels of a main effect term is d. Then, for an interaction involving this main effect, the correlation between the two levels will be d if and only if the factor levels of the other constituent terms are identical. To demonstrate this we return to our pedagogic example where the regions A and B, and B and C are neighbours, but A and C are not neighbours. A CAR structure is specified. In this example, the neighbourhood matrix is

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

so that $\mathbf{D}_t$ for the main effect of region is

$$\mathbf{D}_{\text{region}} = \frac{1}{1-2\tau^2} \begin{pmatrix} 1-\tau^2 & \tau & \tau^2 \\ \tau & 1 & \tau \\ \tau^2 & \tau & 1-\tau^2 \end{pmatrix}.$$

The eigenvalues of $\mathbf{G}$ are $(-\sqrt{2}, 0, \sqrt{2})$, so, therefore, $\tau \in (\tau_{\min}, \tau_{\max}) = (-1/\sqrt{2}, 1/\sqrt{2})$. If an independent correlation structure is specified for the main effect of age, then

$$\mathbf{D}_{\text{age:region}} = \frac{1}{1-2\tau^2} \begin{pmatrix} 1-\tau^2 & \tau & \tau^2 & 0 & 0 & 0 \\ \tau & 1 & \tau & 0 & 0 & 0 \\ \tau^2 & \tau & 1-\tau^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-\tau^2 & \tau & \tau^2 \\ 0 & 0 & 0 & \tau & 1 & \tau \\ 0 & 0 & 0 & \tau^2 & \tau & 1-\tau^2 \end{pmatrix}. \tag{9}$$

The correlation between A and B for the main effect of region is $\tau(1-\tau^2)^{-1/2}$. For the age and region interaction, the correlation between levels involving A and B is $\tau(1-\tau^2)^{-1/2}$ if and only if they have the same level for age. It now follows from (6) and (9) that the scale matrix for the prior distribution is

$$\mathbf{\Sigma}_{\text{age:region}} = \frac{1}{3+4\tau} \begin{pmatrix} 1+\tau & -1/2 \\ -1/2 & 1 \end{pmatrix}.$$

If we denote the regression parameters for this term as $\boldsymbol{\beta}_t = (\beta_{t1}, \beta_{t2})$, where $t = \text{age} : \text{region}$, then the prior correlation between β_{t1} and β_{t2} is

$$\text{corr}\,(\beta_{t1}, \beta_{t2}) = -\frac{1}{2\sqrt{1+\tau}}.$$

If $\tau = 0$, corresponding to independence between the regions, i.e. $\mathbf{D}_t = \mathbf{I}_{q_t}$, and thus we have the Sabanes-Bové and Held [18] prior, then $\text{corr}(\beta_{t1}, \beta_{t2}) = -1/2$. The function $\text{corr}(\beta_{t1}, \beta_{t2})$ is increasing in τ but the correlation is always negative. This is caused by the sum-to-zero constraints. As τ increases, the magnitude of the negative correlation decreases.

A further advantage of using the structure defined by (8) is computational. If we assume that the independence model, containing only the main effect terms, is the simplest model we wish to consider then we will always have the same set of hyperparameters in each model.

3.4. Alternative constraint systems

We now consider alternative constraint systems to sum-to-zero constraints, e.g. corner-point or Helmert constraints. Let $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}$ denote the vectors of regression parameters under the alternative and sum-to-zero constraints, respectively. Since, under the sum-to-zero constraints, each component, $\boldsymbol{\beta}_t$, of $\boldsymbol{\beta}$ has a normal distribution, then $\boldsymbol{\beta}$ has a normal distribution with mean zero and variance matrix $\boldsymbol{\Psi} = \operatorname{diag}\left\{\sigma_1^2\boldsymbol{\Sigma}_1, \ldots, \sigma_T^2\boldsymbol{\Sigma}_T\right\}$. It can be shown (see Appendix B) that

$$\begin{aligned}\boldsymbol{\beta}_A &= \left(\mathbf{X}_A^T\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{X}\boldsymbol{\beta},\\ &= \mathbf{R}_A\mathbf{X}\boldsymbol{\beta},\end{aligned} \tag{10}$$

where $\mathbf{X}_A$ and $\mathbf{X}$ are the model matrices under the alternative and sum-to-zero constraints, respectively, $\mathbf{J}_n$ is the $n \times n$ matrix of ones and

$$\mathbf{R}_A = \left(\mathbf{X}_A^T\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right).$$

Therefore $\boldsymbol{\beta}_A \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Psi}_A\right)$, where the prior variance matrix, $\boldsymbol{\Psi}_A$, is given by

$$\boldsymbol{\Psi}_A = \mathbf{R}_A\mathbf{X}\boldsymbol{\Psi}\mathbf{X}^T\mathbf{R}_A^T.$$

Note that, under the alternative constraints, $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_r$ may no longer, necessarily, be independent. This is equivalent to the fact that $\boldsymbol{\Psi}_A$ (given by the above expression) may no longer, necessarily, be block diagonal.

Under the independence structure described in Section 3.2, where $\mathbf{D}_t = \mathbf{I}_{q_t}$, for $t = 1, \ldots, T$, then

$$\boldsymbol{\Psi}_A = g\mathbf{R}_A\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}_A^T.$$

The matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is called the hat matrix and is invariant to the type of constraint system used, i.e. $\mathbf{H} = \mathbf{H}_A = \mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T$ and therefore

$$\boldsymbol{\Psi}_A = g\left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}.$$

Therefore the proposed prior distribution is a generalisation of the default prior distribution of Sabanes-Bové and Held [18] for any type of constraint system.

4. Example: estimating the number of injecting drug users (IDUs) in Scotland from capture–recapture data

In this section we apply our proposed default prior distribution to an incomplete contingency table which has six factors and 352 cells that involves estimating the number of injecting drug users (IDUs) in Scotland in 2006. These data have been previously analysed by King et al. [12] and Overstall et al. [17]. The six factors are social enquiry reports (2 levels: observed; unobserved); hospital records (2 levels: observed; unobserved); Scottish drug misuse database (2 levels: observed; unobserved); age (2 levels: $\leq$35 years; $>$35 years); gender (2 levels: male; female) and region (11 levels: National Health Service (NHS) board regions—see Fig. 1). The first three factors are sources and the 44 cells which correspond to not being observed by any of these sources for the different age/gender/region combinations have missing counts. Therefore the total population of IDUs, N, is unknown. We use Markov chain Monte Carlo (MCMC) methods to obtain posterior distributions for the missing cell entries and therefore a posterior distribution for the total population of IDUs.

King et al. [12] and Overstall et al. [17] merged the eleven regions into just two levels: Greater Glasgow and Clyde, and the Rest of Scotland. Without merging, using all eleven distinct regions, there are small cell counts for many of the regions. For instance, in one region there are only 19 observed IDUs over all source, age and gender cross-classifications. This suggests that a prior distribution that involves smoothing (or borrowing of information), such as the prior proposed in Section 3, is required.

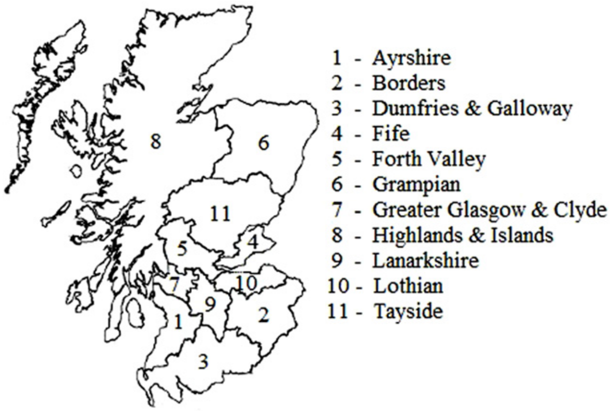

Fig. 1. Map showing the eleven regions of Scotland which correspond to National Health Service (NHS) board regions.

We apply the proposed prior where the independence structure is specified for all of the factors except region where we use the CAR structure described in Section 3.3. By calculating the eigenvalues of the neighbourhood matrix, **G**, for this example, $\tau_{\min} = -0.457$ and $\tau_{\max} = 0.247$. We place a uniform prior on τ in the interval $(\tau_{\min}, \tau_{\max})$. The prior distribution for each $\boldsymbol{\beta}_t$ is

$$\boldsymbol{\beta}_t | \sigma_t^2, \mathbf{D}_t \sim \mathrm{N}\left(0, \sigma_t^2 \boldsymbol{\Sigma}_t\right),$$

where $\boldsymbol{\Sigma}_t$ is given by (6). Following from Section 3.2, we set $\sigma_t^2 = gq_t/n$, with

$$g \sim \mathrm{IG}\left(\frac{a}{2}, \frac{bn}{2}\right),$$

where IG denotes the inverse-gamma distribution, and $a = b = 10^{-3}$, as suggested by Sabanes-Bové and Held [18]. We only specify non-zero prior model probabilities for the log–linear models that contain at most two-way interactions and assume a discrete uniform prior over all of these models. It was found that this allowed enough complexity to obtain an adequate overall model when using the Bayesian p-value to assess model adequacy (see, [8, Chapter 6]).

We use the data-augmentation MCMC approach proposed by King and Brooks [13] with the reversible jump implementation for GLMs of Forster et al. [6] to make moves between log–linear models and the weighted least squares Metropolis–Hastings implementation of Gamerman [7] to make moves within the same log–linear model. We ran the algorithm for one million iterations (discarding the first 10% as burn-in).

For the total population size of IDUs, we obtain a posterior distribution for the total population size with a mean of 21 700 and a 95% highest posterior density interval (HPDI) of (18 900, 24 800). Overstall et al. [17] obtained a posterior mean of 24 000 and a 95% HPDI of (19 500, 29 700) and King et al. [12] a mean of 25 000 with a 95% HPDI of (20 700, 35 000). The advantage of our approach over the latter two analyses is that we are able to provide posterior distributions of the total population size in each NHS board region, broken down by age and gender. Our approach also results in a smaller credible interval for the total population size due to it allowing for correlated regions and not discarding information by merging the factor levels of region.

The posterior mean of τ is 0.108 with a 95% HPDI of $(-0.096, 0.247)$. The posterior probability of τ being positive is 0.816. It follows that the Bayes factor in support of the hypothesis that $\tau > 0$ is 8.205. Therefore there appears to be positive evidence [11] in support of positive spatial correlation between the regions of Scotland.

5. Concluding remarks

In this paper we have proposed a default prior distribution for the regression parameters of a log–linear model that can take account of any dependence structure that may exist between the factor levels. This prior can be applied in situations of model uncertainty and can be seen as a generalisation of other default prior distributions applied to log–linear models including those of Dellaportas and Forster [3], Ntzoufras et al. [15] and Sabanes-Bové and Held [18].

Acknowledgements

The authors thank Dr. Gordon Hay for providing the data in Section 4 and the reviewer for providing helpful comments and suggestions that improved the paper. Both authors were partly funded by MRC-funded addictions cluster, NIQUAD (Grant No. G1000021).

Appendix A. Justification of default prior distribution

In this appendix we give justification for the prior given in Section 3.1, given by (5) and (6). The prior distribution for $\boldsymbol{\beta}_t$ is the conditional distribution of $\boldsymbol{\gamma}^{(1)}$ given that $\boldsymbol{\gamma}^{(2)} = \mathbf{C}_t\boldsymbol{\gamma}^{(1)}$, where $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}\right)^T \sim \mathrm{N}\left(\mathbf{0}, \sigma_t^2\mathbf{M}\right)$, and $\mathbf{M} = \mathbf{P}_t^T\mathbf{D}_t\mathbf{P}_t$. Define

$$\boldsymbol{\psi} = \begin{pmatrix}\boldsymbol{\psi}^{(1)}\\ \boldsymbol{\psi}^{(2)}\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{C}_t & \mathbf{I}\end{pmatrix}\begin{pmatrix}\boldsymbol{\gamma}^{(1)}\\ \boldsymbol{\gamma}^{(2)}\end{pmatrix},$$

so that we now require the conditional distribution of $\boldsymbol{\psi}^{(1)}$ given that $\boldsymbol{\psi}^{(2)} = \mathbf{0}$. It follows, from the properties of the multivariate normal distribution, that $\boldsymbol{\psi}$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma_t^2\mathbf{T}$ where

$$\mathbf{T} = \begin{pmatrix}\mathbf{T}_{11} & \mathbf{T}_{12}\\ \mathbf{T}_{21} & \mathbf{T}_{22}\end{pmatrix},$$

and

$$\begin{aligned}
\mathbf{T}_{11} &= \mathbf{M}_{11},\\
\mathbf{T}_{12} &= \mathbf{M}_{12} - \mathbf{M}_{11}\mathbf{C}_t^T,\\
\mathbf{T}_{21} &= \mathbf{M}_{21} - \mathbf{C}_t\mathbf{M}_{11},\\
\mathbf{T}_{22} &= \mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{C}_t^T\mathbf{C}_t\mathbf{M}_{12} + \mathbf{C}_t\mathbf{M}_{11}\mathbf{C}_t^T.
\end{aligned}$$

The partitioning of $\mathbf{M}$ and $\mathbf{T}$ follows from the partitioning of $\boldsymbol{\gamma}$ into $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$. Using the properties of the multivariate normal distribution the covariance matrix of $\boldsymbol{\beta}_t$ is $\sigma_t^2\boldsymbol{\Sigma}_t$, where

$$\boldsymbol{\Sigma}_t = \mathbf{M}_{11} - \left(\mathbf{M}_{12} - \mathbf{M}_{11}\mathbf{C}_t^T\right)\left(\mathbf{M}_{22} - \mathbf{C}_t\mathbf{M}_{12} - \mathbf{M}_{21}\mathbf{C}_t^T\mathbf{C}_t\mathbf{M}_{11}\mathbf{C}_t^T\right)^{-1}\left(\mathbf{M}_{21} - \mathbf{C}_t\mathbf{M}_{11}\right).$$

Consider the inverse of $\boldsymbol{\Sigma}_t$. It can be shown using, e.g., [10], and after some matrix algebra, that

$$\boldsymbol{\Sigma}_t^{-1} = \mathbf{M}_{11}^{-1} + \mathbf{M}_{11}^{-1}\mathbf{M}_{12}\mathbf{S}_M^{-1}\mathbf{M}_{21}\mathbf{M}_{11}^{-1} - \mathbf{M}_{11}^{-1}\mathbf{M}_{12}\mathbf{S}_M^{-1}\mathbf{C}_t - \mathbf{C}_t^T\mathbf{S}_M^{-1}\mathbf{M}_{21}\mathbf{M}_{11}^{-1} + \mathbf{C}_t^T\mathbf{S}_M^{-1}\mathbf{C}_t,$$

where $\mathbf{S}_M = \mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}$ is the Schur complement (e.g. [9, p. 95]) of $\mathbf{M}_{11}$ in $\mathbf{M}$. As

$$\mathbf{M}^{-1} = \left(\mathbf{P}_t^T\mathbf{D}_t\mathbf{P}_t\right)^{-1} = \mathbf{L} = \begin{pmatrix}\mathbf{L}_{11} & \mathbf{L}_{12}\\ \mathbf{L}_{21} & \mathbf{L}_{22}\end{pmatrix},$$

then it can be shown that

$$\begin{aligned}
\boldsymbol{\Sigma}_t^{-1} &= \mathbf{L}_{11} + \mathbf{L}_{12}\mathbf{C}_t + \mathbf{L}_{21}\mathbf{C}_t^T + \mathbf{C}_t^T\mathbf{L}_{22}\mathbf{C}_t,\\
&= \left(\mathbf{I}\;\mathbf{C}_t^T\right)\left(\mathbf{P}_t^T\mathbf{D}_t^{-1}\mathbf{P}_t\right)\begin{pmatrix}\mathbf{I}\\ \mathbf{C}_t\end{pmatrix},\\
&= \mathbf{A}_t^T\mathbf{D}_t^{-1}\mathbf{A}_t.
\end{aligned}$$

Therefore $\boldsymbol{\Sigma}_t = \left(\mathbf{A}_t^T\mathbf{D}_t^{-1}\mathbf{A}_t\right)^{-1}$ as required.

Appendix B. Correspondence of parameters between different constraint systems

In this appendix we provide a justification of the correspondence between the regression parameters under any constraint system and sum-to-zero constraints, given by (10). Let $\mathbf{Z}_A = (\mathbf{1}_n, \mathbf{X}_A)$ and $\mathbf{Z} = (\mathbf{1}_n, \mathbf{X})$ be the $n \times (p+1)$ matrices formed by appending the vector of ones to the model matrices under the alternative and sum-to-zero constraints. The vector $\left(\phi_A, \boldsymbol{\beta}_A\right)$, where ϕ_A is the intercept under the alternative constraints, is given by

$$\begin{aligned}
\begin{pmatrix} \phi_A \\ \boldsymbol{\beta}_A \end{pmatrix} &= \left(\mathbf{Z}_A^T \mathbf{Z}_A\right)^{-1} \mathbf{Z}_A^T \mathbf{Z} \begin{pmatrix} \phi \\ \boldsymbol{\beta} \end{pmatrix}, \\
&= \begin{pmatrix} n & \mathbf{1}_n^T \mathbf{X}_A \\ \mathbf{X}_A^T \mathbf{1}_n & \mathbf{X}_A^T \mathbf{X}_A \end{pmatrix}^{-1} \begin{pmatrix} n\phi + \mathbf{1}_n^T \mathbf{X}\boldsymbol{\beta} \\ \phi \mathbf{X}_A^T \mathbf{1}_n + \mathbf{X}_A^T \mathbf{X}\boldsymbol{\beta} \end{pmatrix}, \\
&= \begin{pmatrix} \dfrac{1}{n} + \dfrac{1}{n^2} \mathbf{1}_n^T \mathbf{X}_A \mathbf{U}_A^{-1} \mathbf{X}_A^T \mathbf{1}_n & -\dfrac{1}{n} \mathbf{1}_n^T \mathbf{X}_A \mathbf{U}_A^{-1} \\ -\dfrac{1}{n} \mathbf{U}_A^{-1} \mathbf{X}_A^T \mathbf{1}_n & \mathbf{U}_A^{-1} \end{pmatrix} \begin{pmatrix} n\phi + \mathbf{1}_n^T \mathbf{X}\boldsymbol{\beta} \\ \phi \mathbf{X}_A^T \mathbf{1}_n + \mathbf{X}_A^T \mathbf{X}\boldsymbol{\beta} \end{pmatrix},
\end{aligned}$$

where $\mathbf{U}_A = \mathbf{X}_A^T \left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right) \mathbf{X}_A$. The expression for $\boldsymbol{\beta}_A$, given by (10), easily follows.

References

[1] A. Agresti, An Introduction to Categorical Data Analysis, second ed., Wiley, 2007.

[2] N. Cressie, H. Stern, D. Wright, Mapping rates associated with polygons, Journal of Geographical Systems 2 (2000) 61–69.

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

[4] S. Fienberg, The multiple recapture census for closed populations and incomplete 2^k contingency tables, Biometrika 59 (1972) 591–603.

[5] J. Forster, Bayesian inference for Poisson and multinomial log-linear models, Statistical Methodology 7 (2010) 210–224.

[6] J. Forster, R. Gill, A. Overstall, Reversible jump methods for generalised linear models and generalised linear mixed models, Statistics and Computing 22 (2012) 107–120.

[7] D. Gamerman, Sampling from the posterior distribution in generalised linear mixed models, Statistics and Computing 7 (1997) 57–68.

[8] A. Gelman, J. Carlin, H. Stern, D. Rubin, Bayesian Data Analysis, second ed., Chapman and Hall, 2004.

[9] J. Gentle, Matrix Algebra: Theory, Computation, and Applications in Statistics, Springer, 2007.

[10] H. Henderson, S. Searle, On deriving the inverse of a sum of matrices, SIAM Review 23 (1981) 53–60.

[11] R. Kass, A. Raftery, Bayes factors, Journal of the American Statistical Association 90 (1995) 773–795.

[12] R. King, S. Bird, A. Overstall, G. Hay, S. Hutchinson, Injecting drug users in Scotland, 2006: number, demography, and opiate-related death-rates, Addiction Research and Theory 21 (2013) 235–246.

[13] R. King, S. Brooks, On the Bayesian analysis of population size, Biometrika 88 (2001) 317–336.

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

[15] I. Ntzoufras, P. Dellaportas, J. Forster, Bayesian variable and link determination for generalised linear models, Journal of Statistical Planning and Inference 111 (2003) 165–180.

[16] A. O'Hagan, J. Forster, Kendall's Advanced Theory of Statistics, second ed., in: Bayesian Inference, vol. 2B, John Wiley & Sons, 2004.

[17] A. Overstall, R. King, S. Bird, S. Hutchinson, G. Hay, Incomplete contingency tables with censored cells with application to estimating the number of people who inject drugs in Scotland. Tech. Rep., School of Mathematics and Statistics, University of St. Andrews, 2013.

[18] D. Sabanes-Bové, L. Held, Hyper-g priors for generalized linear models, Bayesian Analysis 6 (2011) 387–410.