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
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

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