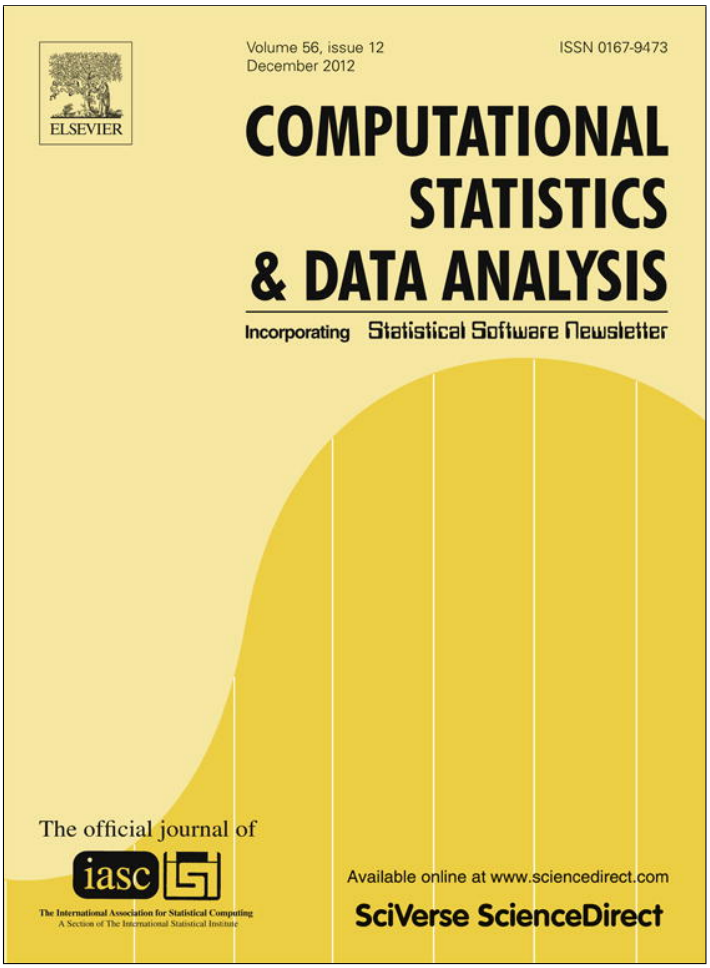

Computational Statistics and Data Analysis 56 (2012) 4067–4080

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data

Francesco Bartolucci [a], Luisa Scaccia [b,*], Alessio Farcomeni [c]

[a] Department of Economics, Finance and Statistics, University of Perugia, Italy

[b] Department of Economic and Financial Institutions, University of Macerata, Italy

[c] Department of Public Health and Infectious Diseases, Sapienza - University of Rome, Italy

ARTICLE INFO

Article history:
Received 7 November 2011
Received in revised form 10 April 2012
Accepted 11 April 2012
Available online 21 April 2012

Keywords:
Bayes factor
Generalized logits
Inequality constraints
Marginal likelihood
Positive association

ABSTRACT

A Bayesian approach is developed for selecting the model that is most supported by the data within a class of marginal models for categorical variables, which are formulated through equality and/or inequality constraints on generalized logits (local, global, continuation, or reverse continuation), generalized log-odds ratios, and similar higher-order interactions. For each constrained model, the prior distribution of the model parameters is specified following the encompassing prior approach. Then, model selection is performed by using Bayes factors estimated through an importance sampling method. The approach is illustrated by three applications based on different datasets, which also include explanatory variables. In connection with one of these examples, a sensitivity analysis to the prior specification is also performed.

1. Introduction

Even if log-linear models are frequently used for the analysis of contingency tables, they do not allow to express, and consequently test, several hypotheses that are usually of interest, mainly because lower order interactions do not refer to the marginal distributions to which they seem to refer. This motivated McCullagh and Nelder (1989, Chapter 6) to introduce a class of models in which the joint distribution of a set of categorical variables is parametrized through the highest log-linear interaction within each possible marginal distribution. Several other models have been proposed following the original idea of McCullagh and Nelder (1989) (see Glonek and McCullagh, 1995; Glonek, 1996; Colombi and Forcina, 2001; Bergsma and Rudas, 2002; Bartolucci et al., 2007).

In this paper, we deal with a flexible class of models in which: (i) the parameters of the saturated model are given by *generalized* logits, in the sense of Douglas et al. (1990), for each univariate marginal distribution, generalized log-odds ratios for each bivariate marginal distribution, and similar interactions for each higher-order marginal distribution; (ii) any constrained model may be formulated through linear equality and inequality constraints on such parameters. In this way, we may express several hypotheses, which are of special interest in the presence of ordinal variables (see Bartolucci et al., 2001; Colombi and Forcina, 2001), as for instance, that: (i) the marginal distribution of one variable is stochastically larger than that of another variable, provided that these have the same categories; (ii) a certain type of positive association between a pair of variables holds; (iii) the marginal distribution of one variable is stochastically increasing with respect to the level of an explanatory variable.

* Correspondence to: Department of Economic and Financial Institutions, University of Macerata, Via Crescimbeni 20, 62100 Macerata, Italy. Tel.: +39 0733 258 3242; fax: +39 0733 258 3205.
E-mail address: scaccia@unimc.it (L. Scaccia).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.04.006

For the above class of models, we develop Bayesian inference and, in particular, a model selection strategy based on the Bayes factor (see Jeffreys, 1935, 1961; Kass and Raftery, 1995), which is defined as the ratio between the marginal likelihoods of two competing models. For the marginal models considered in this paper, the use of the Bayes factor, compared to classical strategies based on the likelihood ratio test, has different advantages. First of all, using the Bayes factor allows for an easy comparison of models parametrized through different types of logit, which would be otherwise cumbersome using a likelihood ratio test. Moreover, since the Bayes factor is computed as the ratio between two marginal likelihoods, the presence of nuisance parameters does not affect the inferential results. Notice that this is a relevant problem in categorical data analysis when studying the association between two variables and the problem also exists when testing for a certain type of positive association using a likelihood ratio test; for a discussion on this point, see Bartolucci et al. (2001) and Dardanoni and Forcina (1998). The distribution of this test statistic depends, in fact, on the nuisance parameters and a possible solution is the conditional approach, which, however, results in a multivariate generalized hypergeometric distribution for the test statistic. Such a distribution is almost intractable whenever the frequencies or the dimension of the table are moderately large, since computing the probability of observing a certain table requires enumerating all the possible tables with the same margins. The strategy based on Monte Carlo maximum likelihood (Bartolucci and Scaccia, 2004) for making inference on the parameters of the model under these constraints helps in overcoming the intractability problem but is still computationally intensive. On the other hand, a drawback of the Bayesian approach proposed here is the need to specify a prior distribution on the parameters, that does not exist within the likelihood ratio approach.

While the decision theoretic approach leads us to select the model with largest marginal likelihood, we can also use the Bayes factor as a measure of evidence. In order to assess this evidence we refer to the Jeffreys (1961) scale, which gives the following guideline: a log Bayes factor below 0.5 indicates *poor* evidence, between 0.5 and 1 *substantial*, between 1 and 2 *strong*, and *decisive* evidence is provided by a log Bayes factor larger than 2. See also Kass and Raftery (1995).

Bayesian methods for the analysis of categorical data have been dealt with by several authors. For instance, Albert (1996, 1997) used the Bayes factor to test hypotheses such as independence, quasi-independence, symmetry, or constant association in two-way and three-way contingency tables. Dellaportas and Forster (1999) proposed a general framework for selecting a log-linear model through the Reversible Jump algorithm of Green (1995) under a multivariate Normal prior distribution on the parameters. In practice, both Albert (1996, 1997) and Dellaportas and Forster (1999) dealt with log-linear models obtained by imposing some linear equality constraints on the parameters of the saturated model; a particular case is the constraint that a subset of the parameters is equal to zero. Klugkist and Hoijtink (2007), Hoijtink et al. (2008), Klugkist et al. (2005a,b, 2010), and Wetzels et al. (2010) used, instead, the Bayes factor to compare competing models expressed through linear inequality and *about equality* constraints on the saturated model. Under their *encompassing prior* approach, the Bayes factor between a constrained model and the encompassing model reduces to the ratio of the probability that the constraints hold under the encompassing posterior distribution and the probability that they hold under the encompassing prior distribution. By encompassing model we mean a model whose parameter space includes that of any other model under consideration. Therefore, once the prior distribution has been specified on the encompassing model parameters (encompassing prior), it is automatically specified for each submodel.

The selection strategy we adopt for the class of models considered in this paper is related to the approach of Klugkist et al. (2010). We exploit their encompassing prior approach, which leads to a logically coherent assessment of prior and posterior model probabilities and parameter distributions, as well as an easy estimation of the Bayes factors. However, our work differs from that of Klugkist et al. (2010) mainly in three respects: (i) we consider a more general class of models for categorical data; (ii) we propose an importance sampling method to improve the efficiency of the Bayes factor estimates for models with very small prior and, possibly, posterior probabilities; (iii) we introduce an iterative algorithm to estimate the Bayes factor for models specified through about equality constraints, which does not require to sample from a constrained model parameter space.

The paper is organized as follows. In Section 2 we describe the class of models of interest. In Section 3 we review the encompassing prior approach and we deal with Bayesian model selection. In Section 4 we illustrate the proposed approach through three applications involving some datasets of interest in the categorical data analysis literature. Finally, we conclude with a brief discussion which is provided in Section 5.

2. Marginal models for categorical variables

In this section, we introduce the class of marginal models developed by McCullagh and Nelder (1989) and illustrate the parametrization based on generalized logits and log-odds ratios; for an overview see also Bergsma et al. (2009). Then, we show how hypotheses of interest may be expressed through linear equality and inequality constraints imposed on the parameters of the saturated model.

2.1. Preliminaries

Let $\boldsymbol{A} = (A_1, \ldots, A_q)$ be a vector of q categorical variables and $\{1, \ldots, m_i\}$ be the support of A_i, $i = 1, \ldots, q$. Also let $r = \prod_i m_i$ be the number of possible configurations of $\boldsymbol{A}$ and let $\boldsymbol{\pi}$ be the r-dimensional column vector of the joint probabilities $\pi_{\boldsymbol{a}} = p(\boldsymbol{A} = \boldsymbol{a})$ arranged in lexicographical order. Suppose, for instance, that there are two categorical variables

($q = 2$) and that the first one has $m_1 = 2$ levels, whereas the second one has $m_2 = 3$ levels. Then, there are $r = 6$ possible configurations of the vector $\boldsymbol{A} = (A_1, A_2)$ and the joint probability vector $\boldsymbol{\pi}$ has the following structure:

$$\boldsymbol{\pi} = \left(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}\right)'. \tag{1}$$

In this context, a possible way to formulate a *log-linear model* (for a general overview about log-linear models, see McCullagh and Nelder, 1989, Chapter 6; Christensen, 1997; Agresti, 2002) is as follows. We first define a vector of log-linear parameters as

$$\boldsymbol{\lambda} = \boldsymbol{H}'\mathbf{log}(\boldsymbol{\pi}), \tag{2}$$

where $\boldsymbol{H}$ is an $r \times (r-1)$ full rank matrix. Note that the above parametrization may be inverted in an explicit way by

$$\boldsymbol{\pi} = \frac{1}{\exp(\boldsymbol{G}\boldsymbol{\lambda})'\mathbf{1}} \exp(\boldsymbol{G}\boldsymbol{\lambda}), \tag{3}$$

where $\boldsymbol{G}$ is an $r \times (r-1)$ full rank matrix such that $\boldsymbol{H}'\boldsymbol{G} = \boldsymbol{I}_{r-1}$, with $\boldsymbol{I}_h$ denoting the identity matrix of dimension h, and $\mathbf{1}$ is a column vector of ones of the appropriate dimension. Then, this parametrization defines a *saturated model*, in the sense that $\boldsymbol{\pi}$, as expressed in (3), may assume any possible value in the simplex of dimension r. Constraints on the log-linear parameters may be expressed as $\boldsymbol{E}\boldsymbol{\lambda} = \mathbf{0}$, for a suitable matrix $\boldsymbol{E}$ and where $\mathbf{0}$ is a column vector of zeros of the appropriate dimension. These concepts are clarified below through a simple example.

Again for the case of two variables with $m_1 = 2$ and $m_2 = 3$ categories, considered above, suppose that

$$\boldsymbol{H}' = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix}.$$

In this way, $\boldsymbol{\lambda}$ contains two conditional logits for the second variable, that is

$$\log \frac{\pi_{12}}{\pi_{11}} = \log \frac{p(A_2 = 2|A_1 = 1)}{p(A_2 = 1|A_1 = 1)} \quad \text{and} \quad \log \frac{\pi_{13}}{\pi_{11}} = \log \frac{p(A_2 = 3|A_1 = 1)}{p(A_2 = 1|A_1 = 1)},$$

one conditional logit for the first variable, that is

$$\log \frac{\pi_{21}}{\pi_{11}} = \log \frac{p(A_1 = 2|A_2 = 1)}{p(A_1 = 1|A_2 = 1)},$$

and the two log-odds ratios

$$\log \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \quad \text{and} \quad \log \frac{\pi_{11}\pi_{23}}{\pi_{13}\pi_{21}}.$$

It is well known that log-odds ratios are measures of association between two categorical variables; this point is discussed in more detail in Section 2.2. In particular, by requiring that all log-odds ratios are equal to 0, we formulate the *model of independence* according to which the joint probabilities are given by $\pi_{\boldsymbol{a}} = \prod_i p(A_i = a_i)$. With reference to the above example, this hypothesis may be formulated as $\boldsymbol{E}\boldsymbol{\lambda} = \mathbf{0}$, with

$$\boldsymbol{E} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4}$$

2.2. Marginal models

In the following, we describe a saturated parametrization of $\boldsymbol{\pi}$ based on marginal logits, marginal log-odds ratios, and similar higher-order interactions.

Marginal logits may be of type *local* (l), *global* (g), *continuation* (c), or *reverse continuation* (r). For the i-th variable, these are defined as follows, for $a_i = 1, \ldots, m_i - 1$:

- *local*: $\eta_i(a_i; l) = \log \dfrac{p(A_i = a_i + 1)}{p(A_i = a_i)}$;
- *global*: $\eta_i(a_i; g) = \log \dfrac{p(A_i \geq a_i + 1)}{p(A_i \leq a_i)}$;

- *continuation*: $\eta_i(a_i; c) = \log \dfrac{p(A_i \geq a_i + 1)}{p(A_i = a_i)}$;
- *reverse continuation*: $\eta_i(a_i; r) = \log \dfrac{p(A_i = a_i + 1)}{p(A_i \leq a_i)}$.

Local logits are used when it is of interest to compare the marginal probability of each category with that of the previous category. Logits of type global and continuation are specially tailored to deal with ordinal variables. In particular, logits of type global are more appropriate when the variable may be seen as a discretized version of an underlying continuum, whereas logits of type continuation are more appropriate when categories correspond to levels of achievement that may be entered only if the previous level has also been achieved, as in education. Finally, using logits of type reverse continuation is the same as arranging categories in reverse order and using logits of type continuation.

Marginal log-odds ratios are defined as contrasts between conditional logits. For two variables, A_i and A_j, the most well-known log-odds ratios are shown in the following, where $a_i = 1, \ldots, m_i - 1$ and $a_j = 1, \ldots, m_j - 1$:

- *Local*: when logits of type l are used for both A_i and A_j

$$\begin{aligned}\eta_{ij}(a_i, a_j; l, l) &= \eta_j(a_j; l|A_i = a_i + 1) - \eta_j(a_j; l|A_i = a_i)\\ &= \log \frac{p(A_i = a_i, A_j = a_j)p(A_i = a_i + 1, A_j = a_j + 1)}{p(A_i = a_i, A_j = a_j + 1)p(A_i = a_i + 1, A_j = a_j)};\end{aligned}$$

- *Local–Global*: when logits of type l are used for A_i and of type g for A_j

$$\begin{aligned}\eta_{ij}(a_i, a_j; l, g) &= \eta_j(a_j; g|A_i = a_i + 1) - \eta_j(a_j; g|A_i = a_i)\\ &= \log \frac{p(A_i = a_i, A_j \leq a_j)p(A_i = a_i + 1, A_j \geq a_j + 1)}{p(A_i = a_i, A_j \geq a_j + 1)p(A_i = a_i + 1, A_j \leq a_j)};\end{aligned}$$

- *Global*: when logits of type g are used for both A_i and A_j

$$\begin{aligned}\eta_{ij}(a_i, a_j; g, g) &= \eta_j(a_j; g|A_i \geq a_i + 1) - \eta_j(a_j; g|A_i \leq a_i)\\ &= \log \frac{p(A_i \leq a_i, A_j \leq a_j)p(A_i \geq a_i + 1, A_j \geq a_j + 1)}{p(A_i \leq a_i, A_j \geq a_j + 1)p(A_i \geq a_i + 1, A_j \leq a_j)}.\end{aligned}$$

Similarly, three-way interactions are defined as contrasts between conditional log-odds ratios and so on for higher order interactions.

As mentioned in Section 2.1, marginal parameters cannot be directly expressed by a log-linear parametrization of type (2). To clarify this point, consider again the case of two variables for which the joint probability vector has the structure in (1). Then, for instance, we may formulate a model based on the global logits for the second variable, that is

$$\begin{aligned}\eta_2(1; g) &= \log \frac{p(A_2 \geq 2)}{p(A_2 = 1)} = \log \frac{\pi_{12} + \pi_{13} + \pi_{22} + \pi_{23}}{\pi_{11} + \pi_{21}},\\ \eta_2(2; g) &= \log \frac{p(A_2 = 3)}{p(A_2 \leq 2)} = \log \frac{\pi_{13} + \pi_{23}}{\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22}}.\end{aligned}$$

Moreover, we may model the association between the two variables through the following two local–global log-odds ratios:

$$\begin{aligned}\eta_{12}(1, 1; l, g) &= \log \frac{p(A_1 = 1, A_2 = 1)p(A_1 = 2, A_2 \geq 2)}{p(A_1 = 1, A_2 \geq 2)p(A_1 = 2, A_2 = 1)} = \log \frac{\pi_{11}(\pi_{22} + \pi_{23})}{(\pi_{12} + \pi_{13})\pi_{21}},\\ \eta_{12}(1, 2; l, g) &= \log \frac{p(A_1 = 1, A_2 \leq 2)p(A_1 = 2, A_2 = 3)}{p(A_1 = 1, A_2 = 3)p(A_1 = 2, A_2 \leq 2)} = \log \frac{(\pi_{11} + \pi_{12})\pi_{23}}{\pi_{13}(\pi_{21} + \pi_{22})}.\end{aligned}$$

The previous expressions clarify that these types of logit and log-odds ratio cannot be expressed through the log-linear form in (2), since they involve logarithms of sums of suitable joint probabilities. As we illustrate in Section 2.3, these effects are necessary to express certain hypotheses of interest, such as that of marginal homogeneity or certain forms of positive dependence. On the other hand, marginal models are based on a parametrization that allows us to include these effects. This is explained in the following.

Let $\boldsymbol{z}$ be a q-dimensional vector of zeros and ones, let $\boldsymbol{\eta}_{\boldsymbol{z}}$ be a column vector containing all the marginal interactions between the variables A_i such that $z_i = 1$ and let $\boldsymbol{\eta}$ be the vector obtained by stacking, in lexicographical order, the vectors $\boldsymbol{\eta}_{\boldsymbol{z}}$, for all $\boldsymbol{z} \neq \boldsymbol{0}$. Following Colombi and Forcina (2001), such a vector, which provides the saturated parametrization of $\boldsymbol{\pi}$ at issue, may be simply obtained through the following expression, which may be seen as an extension of (2):

$$\boldsymbol{\eta} = \boldsymbol{C} \log(\boldsymbol{M}\boldsymbol{\pi}), \tag{5}$$

where $\boldsymbol{C}$ and $\boldsymbol{M}$ are appropriate matrices, whose construction is described in Appendix A; see also Lang (1996). However, to invert Eq. (5), and so obtain $\boldsymbol{\pi}$ in terms of $\boldsymbol{\eta}$, we must rely on a Newton algorithm as the one described in Glonek and McCullagh (1995) and Colombi and Forcina (2001); this inversion method is also illustrated in Appendix A. The number of

elements in $\boldsymbol{\eta}$ is equal to $r-1$, which is the number of free elements of the joint probability vector $\boldsymbol{\pi}$, taking into account the constraint $\boldsymbol{\pi}'\mathbf{1}=1$.

In order to clarify the parametrization in (5), consider again the example above based on two categorical variables. In this case, the vector of marginal parameters has the following structure

$$\boldsymbol{\eta}=\left(\eta_2(1;g),\eta_2(2;g),\eta_1(1;l),\eta_{12}(1,1;l,g),\eta_{12}(1,2;l,g)\right)', \tag{6}$$

which is obtained by letting

$$\boldsymbol{C}=\begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \end{pmatrix},\qquad \boldsymbol{M}=\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0\\ 1 & 1 & 0 & 1 & 1 & 0\\ 0 & 1 & 1 & 0 & 1 & 1\\ 0 & 0 & 1 & 0 & 0 & 1\\ 1 & 1 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 1 & 1\\ 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 1 & 0\\ 0 & 0 & 0 & 0 & 1 & 1\\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that, in this example, the parameter vector $\boldsymbol{\eta}$ defined in (6) has dimension 5, which is equal to the number of free elements of the corresponding joint probability vector $\boldsymbol{\pi}$ defined in (1).

2.3. Constrained models

A variety of constrained models may be formulated by posing linear equality and inequality constraints of the form

$$\boldsymbol{E}\boldsymbol{\eta}=\mathbf{0},\qquad \boldsymbol{U}\boldsymbol{\eta}\geq\mathbf{0}, \tag{7}$$

on the saturated parameter vector. Here, and throughout the paper, equality constraints are substituted by about equality constraints of type $|\boldsymbol{E}\boldsymbol{\eta}|\leq\boldsymbol{\epsilon}$, for a vector $\boldsymbol{\epsilon}>\mathbf{0}$ having suitably small elements; this choice is motivated in Section 3.3.

Consider first the case of only two variables, A_1 and A_2. The most interesting hypotheses are usually on the association between these variables. Let $c=m_1+m_2-2$ be the number of the marginal logits and $d=(m_1-1)(m_2-1)$ be that of the log-odds ratios. For instance, in the case of two variables A_1 and A_2 with 2 and 3 categories, respectively, we have $m_1-1=1$ marginal logits for A_1, $m_2-1=2$ marginal logits for A_2, so that $c=3$, and $d=2$ marginal log-odds ratios.

By requiring that all the log-odds ratios are non-negative, namely by letting

$$\boldsymbol{U}=\begin{pmatrix}\boldsymbol{O}_{d,c} & \boldsymbol{I}_d\end{pmatrix}, \tag{8}$$

we express the hypothesis of positive association between A_1 and A_2, where, in general, $\boldsymbol{O}_{h,j}$ is a matrix of $h\times j$ zeros. For instance, when $m_1=2$ and $m_2=3$, $\boldsymbol{U}$ is equal to the matrix given in (4). Obviously, the type of association depends on the type of logit that is used for the two variables. For instance, with local logits for both variables we are formulating the hypothesis of *Total Positivity of Order* 2 (TP_2; see Karlin, 1968), whereas with global logits for both variables we are formulating the hypothesis of *Positive Quadrant Dependence* (PQD; see Lehmann, 1966). Note that there is a hierarchy among these notions of positive association in the sense that, for instance, TP_2 implies that all the continuation log-odds ratios are non-negative which, in turn, implies PQD (see Douglas et al., 1990, for details). Also note that, regardless of the type of log-odds ratio, independence between A_1 and A_2 may be expressed through the constraint that $\boldsymbol{E}$, rather than $\boldsymbol{U}$, is equal to the matrix in (8). A less stringent constraint than that of independence is the constraint of uniform association, namely that all the d log-odds ratios are equal to each other (Plackett, 1965). This type of constraint is formulated by letting

$$\boldsymbol{E}=\begin{pmatrix}\boldsymbol{O}_{d-1,c} & \boldsymbol{D}_{d-1}\end{pmatrix},$$

where, in general, $\boldsymbol{D}_h=\begin{pmatrix}\mathbf{0} & \boldsymbol{I}_{h-1}\end{pmatrix}-\begin{pmatrix}\boldsymbol{I}_{h-1} & \mathbf{0}\end{pmatrix}$ is a matrix that produces first differences.

When A_1 and A_2 have the same categories, the number of which is indicated by m, constraints on the univariate marginal distributions may also be of interest. For instance, we may formulate the constraint of marginal homogeneity by letting

$$\boldsymbol{E}=\begin{pmatrix}-\boldsymbol{I}_{m-1} & \boldsymbol{I}_{m-1} & \boldsymbol{O}_{m-1,d}\end{pmatrix}.$$

When global logits are used for both variables, the constraint that A_2 is stochastically greater than A_1 may be imposed by letting $\boldsymbol{U}$, rather than $\boldsymbol{E}$, equal to the matrix above. When there are more than two variables, similar constraints may also involve interactions of order higher than two. These are typically of interest in the presence of longitudinal data. Some useful insights on how formulating a marginal model in these situations are provided by Glonek and McCullagh (1995).

The approach outlined above may be easily extended when dealing with one or more explanatory variables, which are collected in the vector $\boldsymbol{B}$. Every possible configuration of these variables, say $\boldsymbol{b}$, defines a stratum conditionally on which we have a vector of joint probabilities of the response variables, $\boldsymbol{\pi}(\boldsymbol{b})$, and a vector of marginal parameters, $\boldsymbol{\eta}(\boldsymbol{b})$, defined as in (5). Obviously, in this setting we may impose the same constraints illustrated above within each stratum and also constraints involving the parameters of different strata. These constraints are still of the type $\boldsymbol{E}\boldsymbol{\eta} = \boldsymbol{0}$, $\boldsymbol{U}\boldsymbol{\eta} \geq \boldsymbol{0}$, but in this case by $\boldsymbol{\eta}$ we mean the vector obtained by stacking the vectors $\boldsymbol{\eta}(\boldsymbol{b})$ for every $\boldsymbol{b}$. For instance, when we only have two response variables, we may express the constraint of conditional independence between A_1 and A_2, given $\boldsymbol{B}$, by

$$\boldsymbol{E} = \boldsymbol{I}_s \otimes \begin{pmatrix} \boldsymbol{O}_{d,c} & \boldsymbol{I}_d \end{pmatrix},$$

where s is the number of strata, that is the number of different configurations of $\boldsymbol{B}$, and $\otimes$ denotes the Kronecker product. For instance, if $q = 2$ and, as in the example considered so far, $m_1 = 2$ and $m_2 = 3$, with $s = 2$ strata we have

$$\boldsymbol{E} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We may also formulate the hypothesis that A_1 and A_2 have the same degree of association for each stratum by letting

$$\boldsymbol{E} = \boldsymbol{D}_s \otimes \begin{pmatrix} \boldsymbol{O}_{d,c} & \boldsymbol{I}_d \end{pmatrix},$$

which, for the example at hand, becomes

$$\boldsymbol{E} = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

or the hypothesis that the explanatory variables do not affect the marginal distributions of A_1 and A_2, by letting

$$\boldsymbol{E} = \boldsymbol{D}_s \otimes \begin{pmatrix} \boldsymbol{I}_c & \boldsymbol{O}_{c,d} \end{pmatrix}.$$

Finally, if we have only one explanatory variable, B, and this is ordinal, we can express the constraint that the marginal distributions of A_1 and A_2 increase with the level of B by letting $\boldsymbol{U}$, rather than $\boldsymbol{E}$, equal to the matrix above.

3. Bayesian estimation and model selection

In this section we show how to make inference on the models presented in the previous section. In particular, Section 3.1 is devoted to the choice of appropriate priors for the class of models at issue, Section 3.2 is focused on assessing the plausibility of the different models, given an observed contingency table, Section 3.3 deals with computational issues which arise in estimating the Bayes factor, and Section 3.4 illustrates how to perform parameter estimation in an efficient way.

In the following, when the data are not stratified, we denote the frequency corresponding to the configuration $\boldsymbol{a}$ of the observed table by $y_{\boldsymbol{a}}$ and by $\boldsymbol{y}$ the vector with elements $y_{\boldsymbol{a}}$ arranged as in $\boldsymbol{\pi}$, so that, with $q = 2$, $m_1 = 2$, and $m_2 = 3$, for example, we have $\boldsymbol{y} = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23})'$. When the data are stratified according to one or more explanatory categorical variables, we have a vector of frequencies $\boldsymbol{y}(\boldsymbol{b})$ for every configuration of such variables and, consequently, $\boldsymbol{y}$ denotes the vector obtained by stacking the vectors $\boldsymbol{y}(\boldsymbol{b})$ for every $\boldsymbol{b}$.

3.1. Prior distributions

In a Bayesian framework, it is natural to include equality and inequality constraints imposed on the model parameters as prior knowledge. Since all the constrained models presented in Section 2.3 are nested in an unconstrained (or encompassing) model, we use the concept of encompassing prior (Klugkist et al., 2005a,b, 2010; Klugkist and Hoijtink, 2007). Therefore, we specify the prior distribution only for an encompassing model and then we derive the prior distributions for the other models by restricting the parameter space according to the constraints of interest. This approach has a very nice interpretation: the resulting Bayes factor for model selection (see Section 3.2) coincides with the ratio between the proportions of the parameter space that are in agreement with the constrained model, under the posterior and the prior distributions of the encompassing model. This approach also has the advantage that only one single prior distribution needs to be specified. Moreover, the method can be seen as a generalization of the method based on the Savage–Dickey density ratio, which overcomes the Borel–Kolmogorov paradox (Dawid and Lauritzen, 2001). See Wetzels et al. (2010) for a detailed discussion on this point.

In the present framework, it is natural to choose the saturated model based on the parameter vector $\boldsymbol{\pi}$ as the encompassing model. Under this model, the parameter space is the simplex of dimension r and the frequency vector $\boldsymbol{y}$ has multinomial distribution with parameters n, the number of observations, and $\boldsymbol{\pi}$. The choice of the $\boldsymbol{\pi}$ parametrization, rather than the parametrization based on the vector of marginal parameters $\boldsymbol{\eta}$, is motivated by the fact that it also makes straightforward the comparison between different types of logit.

Let M_1 indicate the saturated (encompassing) model and let $p(\boldsymbol{\pi}|M_1)$ denote the encompassing prior distribution. The prior distribution of each constrained model M_k, for $k = 2, \ldots, K$, directly follows from this prior as

$$p(\boldsymbol{\pi}|M_k) = \frac{p(\boldsymbol{\pi}|M_1)\delta_k(\boldsymbol{\pi})}{\int p(\boldsymbol{\pi}|M_1)\delta_k(\boldsymbol{\pi})\mathrm{d}\boldsymbol{\pi}} = \alpha_k p(\boldsymbol{\pi}|M_1)\delta_k(\boldsymbol{\pi}), \tag{9}$$

where the integral is on the simplex of dimension r. Moreover, $\delta_k(\boldsymbol{\pi})$ is the indicator function equal to 1 if $\boldsymbol{\pi}$ is in accordance with the constraints defining model M_k and to 0 otherwise and α_k is the inverse of the proportion of the parameter space that, under the encompassing prior, is in agreement with these constraints. Obviously, the constrained prior in (9) is not defined for a model with equality constraints, but it is defined for a model formulated by about equality constraints.

Under the encompassing prior approach, the posterior distribution of the parameters for each constrained model also follows immediately from the posterior under the encompassing model. In particular, we have

$$p(\boldsymbol{\pi}|\boldsymbol{y}, M_k) = \frac{p(\boldsymbol{\pi}|\boldsymbol{y}, M_1)\delta_k(\boldsymbol{\pi})}{\int p(\boldsymbol{\pi}|\boldsymbol{y}, M_1)\delta_k(\boldsymbol{\pi})\mathrm{d}\boldsymbol{\pi}} = \beta_k p(\boldsymbol{\pi}|\boldsymbol{y}, M_1)\delta_k(\boldsymbol{\pi}), \tag{10}$$

where, now, β_k is the inverse of the proportion of the parameter space that, under the encompassing posterior, is in agreement with the constraints of model M_k.

Coming to the issue of choosing the encompassing prior $p(\boldsymbol{\pi}|M_1)$ for $\boldsymbol{\pi}$, the uniform distribution on the simplex of dimension r has been acknowledged to be the default prior. This is equivalent to assuming for model M_1 that, *a priori*, $\boldsymbol{\pi} \sim D(\mathbf{1})$, where $D(\mathbf{1})$ denotes the Dirichlet distribution with parameter $\mathbf{1}$. See, for instance, Tuyl et al. (2009) for a detailed discussion on this choice.

The posterior for the saturated parametrization $\boldsymbol{\pi}$, with the default prior choice, is readily derived and it is of type $D(\mathbf{1}_r + \boldsymbol{y})$. Therefore, samples can be independently drawn from the prior and posterior distributions for the saturated model and the corresponding normalizing constants are available in closed form.

3.2. Model selection

Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ denote the set of models of interest. As already noted, with the exception of the saturated model M_1, each of these models is defined by a certain type of logit for every response variable and by constraints of type (7) on the vector of marginal parameters. Models $M_2, \ldots, M_K$ are all nested in M_1, but not necessarily nested in one another.

For model selection, we make use of the *Bayes factor*, which is the ratio of the marginal likelihoods of two competing models. Thus, the Bayes factor for model M_k versus the encompassing model is defined as:

$$B_{k1} = \frac{p(\boldsymbol{y}|M_k)}{p(\boldsymbol{y}|M_1)} = \frac{\int p(\boldsymbol{y}|\boldsymbol{\pi}, M_k)p(\boldsymbol{\pi}|M_k)\mathrm{d}\boldsymbol{\pi}}{\int p(\boldsymbol{y}|\boldsymbol{\pi}, M_1)p(\boldsymbol{\pi}|M_1)\mathrm{d}\boldsymbol{\pi}},$$

where $p(\boldsymbol{y}|\boldsymbol{\pi}, M_k)$ and $p(\boldsymbol{y}|M_k)$ denote the likelihood of the data and the marginal likelihood for model M_k, respectively. The Bayes factor measures the evidence that the data provide for one model versus the other and corresponds to the change from prior model odds to posterior model odds. Obviously, the larger is B_{k1}, the greater is the evidence provided by the data in favor of M_k with respect to M_1. So, when B_{k1} is larger than 1, or equivalently $\log(B_{k1}) > 0$, model M_k has to be preferred to model M_1. To compare more than two models, or equivalently to choose the best model in $\mathcal{M}$ when $K > 2$, a convenient possibility is to single out M_1 as the reference model and then to compute the Bayes factor between any other model and the unconstrained one, that is B_{k1}, for $k = 2, \ldots, K$. The model to be preferred is that with the largest Bayes factor, provided that it is larger than 1; otherwise the best model is the saturated model. Obviously, the Bayes factor for comparing every pair of models M_k and M_l, which are not necessarily nested, is straightforwardly computed as $B_{kl} = B_{k1}/B_{l1}$.

It is important to note that the Bayes factor, as model selection tool, combines goodness of fit with a correction for model complexity; see, among others, Spiegelhalter and Smith (1982), Jefferys and Berger (1992), Kass and Raftery (1995) and the references therein.

3.3. Computational issues in estimating the Bayes factor

Direct computation of the Bayes factor is almost always infeasible, and this also happens for the class of models dealt with here. Several methods have been proposed to estimate the Bayes factor numerically, but the estimation is generally cumbersome from the computational point of view (see Gamerman and Lopes, 2006, Chapter 7, for a review).

The encompassing prior approach renders a nice interpretation of the Bayes factor for a constrained model M_k with respect to the encompassing model M_1, which virtually eliminates the computational complications in the Bayes factor estimation. In fact, as demonstrated in Klugkist et al. (2005a), the Bayes factor for a constrained model M_k versus the encompassing model M_1 reduces to the ratio between the proportions of the parameter space that are in agreement with the constraints defining M_k under the posterior and under the prior distribution of M_1, that is

$$B_{k1} = \frac{\alpha_k}{\beta_k}. \tag{11}$$

The above result implies that estimating the Bayes factor is particularly simple. The encompassing prior is sampled and α_k is estimated as $\hat{\alpha}_k$, which is the inverse of the proportion of the sample draws that are in agreement with the constraints defining model M_k. Similarly, sampling from the encompassing posterior allows us to estimate β_k as $\hat{\beta}_k$, which is the inverse of the proportion of the draws that are in agreement with the constraints of model M_k. In this way, using just one sample of a large number of draws from the encompassing prior and another from the encompassing posterior, the estimate

$$\hat{B}_{k1} = \frac{\hat{\alpha}_k}{\hat{\beta}_k}$$

can be computed for each constrained model M_k, $k = 2, \ldots, K$.

Note that, in our setting, the choice of the Dirichlet default prior for $\boldsymbol{\pi}$ allows to sample independently under both the encompassing prior and posterior distributions, leading to further simplifications in estimating the Bayes factor. Moreover, in some cases α_k can be exactly computed without the need of sampling from the encompassing prior. However, there are two issues that we must deal with when estimating the Bayes factor and that we illustrate in the following.

First of all, a rare event problem can arise. Consider for instance our example in Section 4.1: we have a six by six contingency table with 35 free parameters under the unconstrained model. The hypothesis of positive association is formulated by requiring the positivity of the 25 log-odds ratios. When using logits of type l for both variables, the constant $1/\alpha_k$ for a positive association model can be calculated exactly as $0.5^{25} = 2.9802 \times 10^{-8}$. In this case, sampling from the encompassing prior is not required, but such a small value of $1/\alpha_k$ can be common to other models. For these models, even if we drew millions of values from the encompassing prior, we would expect to draw no values satisfying the constraint. This would lead an estimate of the Bayes factor equal to ∞ or to ∞/∞, in case the same problem also arises when sampling from the encompassing posterior. Furthermore, even if a finite estimate of the Bayes factor can be obtained, its variance would be huge.

The problem described above is that of *rare event simulation* (e.g., Bucklew, 2004), which is often overcome through *importance sampling*. Suppose we want to estimate $1/\alpha_k$. From (9) it immediately follows that $1/\alpha_k = E_p[\delta_k(\boldsymbol{\pi})]$, where the expected value is calculated with respect to the encompassing prior $p(\boldsymbol{\pi}|M_1)$. Now, letting $g(\boldsymbol{\pi})$ be any density such that $p(\boldsymbol{\pi}|M_1) = 0$ whenever $g(\boldsymbol{\pi}) = 0$, we can re-write

$$E_p[\delta_k(\boldsymbol{\pi})] = \int p(\boldsymbol{\pi}|M_1)\delta_k(\boldsymbol{\pi})\mathrm{d}\boldsymbol{\pi} = \int \left[\frac{p(\boldsymbol{\pi}|M_1)}{g(\boldsymbol{\pi})}\delta_k(\boldsymbol{\pi})\right] g(\boldsymbol{\pi})\mathrm{d}\boldsymbol{\pi} = E_g\left[\delta_k(\boldsymbol{\pi})\frac{p(\boldsymbol{\pi}|M_1)}{g(\boldsymbol{\pi})}\right], \tag{12}$$

where the last expected value is now calculated under the *importance* density $g(\boldsymbol{\pi})$. Then, an importance sampling estimate of $1/\alpha_k$ can be obtained by sampling $\boldsymbol{\pi}$ from an appropriate importance density and estimating the last expected value in (12) through the sample mean. If required, an estimate of $1/\beta_k$ can be obtained in a similar way, after choosing an appropriate sampling density.

In this paper, we propose an automatic way to obtain an adequate importance sampling density. Suppose we want to estimate $1/\beta_k$ for a certain model M_k; then the proposed method is based on the following steps:

(i) compute the maximum likelihood estimate of the vector $\boldsymbol{\eta}$ under the constraints imposed by model M_k (see Colombi and Forcina, 2001) and indicate this estimate by $\hat{\boldsymbol{\eta}}$;
(ii) convert $\hat{\boldsymbol{\eta}}$ into $\hat{\boldsymbol{\pi}}$ using the Newton algorithm described in Appendix A; see also Glonek and McCullagh (1995) and Colombi and Forcina (2001);
(iii) choose as importance density a Dirichlet distribution with mean vector equal to $\hat{\boldsymbol{\pi}}$, that is $g(\boldsymbol{\pi}) \sim D(\phi\hat{\boldsymbol{\pi}})$, where ϕ is a tuning parameter that can be appropriately chosen so that enough draws from the importance density satisfy the constraints imposed by model M_k. The optimal tuning parameter could be chosen by minimizing the variance of the approximation, but this expression depends itself on the target quantity. A simple approach, which we use in this paper, is to try different values of ϕ on a suitable grid (say from 0.02 to 50).

The same strategy can be adopted for choosing an appropriate importance density to estimate $1/\alpha_k$. In this case, the maximum likelihood estimate $\hat{\boldsymbol{\eta}}$ in (i) will be that corresponding to a hypothetical contingency table having a vector of frequencies $\boldsymbol{y}$ with all elements equal to a constant value.

A way to assess the success of the importance sampling is given by computation of Effective Sample Size (ESS); see, among others, Liu (2001). Let

$$w(\boldsymbol{\pi}) = \frac{p(\boldsymbol{\pi}|M_1)}{g(\boldsymbol{\pi})}$$

denote the unnormalized importance weight in (12), and v denote the number of draws from $g(\boldsymbol{\pi})$. The ESS is defined as

$$\mathrm{ESS}(v) = \frac{v}{1 + \mathrm{cv}^2}, \tag{13}$$

where cv denotes the coefficient of variation of the weights. A possible general strategy for tuning the proposal distribution is by choosing a tuning parameter giving an ESS not too smaller than v.

To give an idea of the precision of the algorithm, we consider again the example in Section 4.1, to which we already referred. For those data, we compared the true value of $1/\alpha_k$, exactly computable for the TP_2 model, with its estimates

Table 1
True and estimated $1/\alpha_k$ for the model including TP_2 on the basis of the data in Section 4.1.

True value	Estimate #1	Estimate #2	Estimate #3
2.9802×10^{-8}	2.2315×10^{-8}	2.8510×10^{-8}	3.5776×10^{-8}

obtained in three separate runs of the algorithm. The results, which are given in Table 1, show that the approximation is rather satisfactory in all cases. A more convincing evidence was also obtained by independently repeating sampling 100 times. For 93% of these repetitions, the true value of $1/\alpha_k$ fell in the 95% confidence interval calculated around the estimated value $1/\hat{\alpha}_k$.

The second issue in estimating the Bayes factor arises in the presence of equality constraints. For models formulated through these constraints, the Bayes factor cannot be interpreted as the ratio between the proportions of encompassing posterior and prior in agreement with the constraints, since these proportions would be exactly zero. However, it has been recently shown (Wetzels et al., 2010) that the encompassing approach naturally extends to exact equality constraints by considering the ratio of the heights for the encompassing posterior and prior distributions evaluated under the constraint (i.e., the Savage–Dickey density ratio). However, this approach to handle hypotheses specified through exact equality constraints complicates the computation of the Bayes factor for models containing both equality and inequality constraints. For this reason, we rather prefer to follow the idea of Berger and Sellke (1987) and Klugkist et al. (2010) of substituting exact equality with about equality constraints. In this way, the interpretation of the Bayes factor provided in (11) is preserved and models containing inequality or about equality constraints, as well as a mix of both constraints, can be handled in a unified manner. Moreover, Berger and Delampady (1987) noted that a Bayes factor based on equality constraints is indistinguishable from a Bayes factor based on about equality constraints, provided that the interval around the exact equality constraint is small enough. However, if this interval is too small, we incur again in the rare event problem illustrated above, when trying to estimate $1/\alpha_k$ and $1/\beta_k$.

To solve the above problem, Klugkist et al. (2010) proposed a stepwise procedure which guarantees that a small enough interval is used and does not actually need to pre-specify the size of this interval. In principle, we could use this method to estimate the required constants α_k and β_k. This method is based on drawing random numbers from suitably truncated Gamma distributions, which are then normalized to obtain the vector $\boldsymbol{\pi}$. The way in which the support of these distributions is chosen depends on the adopted constraints. In our case, however, the complexity of the models implies that it is difficult to define how the support of each of these variables must be constrained; on the other hand, a rejection sampling procedure to draw random values from the truncated normal would be rather slow. For these reasons, we prefer to adapt the iterative procedure in Klugkist et al. (2010) employing again the importance sampling method. According to our procedure, only two different samples, one drawn from the importance density for the prior and the other one from the importance density for the posterior, are required to estimate the Bayes factor, thus overcoming the problem of sampling from constrained distributions, which affects the procedure in Klugkist et al. (2010). The details of the proposed algorithm are given in Appendix B.

As a further remark, we stress that the complexity of the constraints imposed by the models considered here makes unfeasible, or at least very time consuming, to estimate the marginal likelihood of each constrained model using, for example, the approach in Chib (1995), since this would require an MCMC algorithm for sampling from the constrained posterior distribution of the parameters. Using the parametrization in $\boldsymbol{\eta}$ would allow to exploit the approach in Chib (1995) but would complicate the comparison between models based on different types of logit.

3.4. Parameter estimation

Coming to the issue of parameter estimation, we have to recall that, for the class of models considered here, obtaining point or interval estimates of the parameters is not in general of great interest. The main interest is rather on model selection as a tool for evaluating which hypothesis is mostly supported by the data. Nevertheless, once a particular model M_k has been selected for the data at hand, Bayesian parameter estimation is based on the posterior distribution of the parameters and, in our setting, a sample from this posterior is already available after model choice, as a byproduct of the procedure to estimate β_k. In particular, we take the set of all the draws from the parameter posterior distribution of the saturated model, $D(\mathbf{1}+\boldsymbol{y})$, that are in agreement with the constraints imposed by model M_k. This set should contain enough draws to be also used for parameter estimation purposes.

Obviously, parameter estimates can be obtained in the way described above for models defined by about equality constraints but not for models defined by exact equality constraints. As an alternative, if estimates under exact equality constraints are required, the parametrization in $\boldsymbol{\eta}$ can be used, after choosing an appropriate prior distribution on this parameter vector, for example a Gaussian distribution. However, as already noticed, such a parametrization would complicate model selection in the presence of models expressed through different types of logit.

Table 2
Father (A_1) and son (A_2) occupational status for a sample of 3488 British males.

A_1	A_2					
	I	II	III	IV	V	VI
I	125	60	26	49	14	5
II	47	65	66	123	23	21
III	31	58	110	223	64	32
IV	50	114	185	715	258	189
V	6	19	40	179	143	71
VI	3	14	32	141	91	106

4. Applications

In the following, we illustrate the proposed approach through three applications involving some interesting datasets which also include explanatory variables. In the first application, illustrated in Section 4.1, we also show the results of a sensitivity analysis with respect to prior specification.

4.1. Classification of men by social class and social class of their fathers

We first consider a dataset (see Table 2) referred to a sample of British males cross-classified according to their occupational status (A_2) and that of their father (A_1). The data have been already analyzed by other authors, such as Goodman (1991) and Dardanoni and Forcina (1998). In particular, Dardanoni and Forcina (1998), following a likelihood ratio approach, concluded that the data conform to some forms of positive association. However, due to the presence of nuisance parameters, given by marginal column probabilities, they did not reach a definitive conclusion about TP_2.

For these data we first compared the saturated model (M_1) with the independence model (M_2), the saturated model incorporating PQD (M_3) and that incorporating TP_2 (M_4). For each of the latter three models we estimated the Bayes factor with respect to the saturated model, taken as reference model, through the algorithm described in Section 3.2. We obtained the following results:

$\log(\hat{B}_{21})$	$\log(\hat{B}_{31})$	$\log(\hat{B}_{41})$
−34.88	4.32	5.12

To compute $\hat{B}_{21}$ (see Section 3.3) we used $\phi = 20$ for the prior and $\phi = 1$ for the posterior, where as importance distribution we used a Dirichlet distribution with parameters corresponding to the maximum likelihood estimate of the parameters under the independence model. Since we have about equality constraints, we used the algorithm in Appendix B starting from $\boldsymbol{\epsilon} = 0.1 \cdot \mathbf{1}$, with tuning parameter $b = 0.5$. The algorithm was stopped after two iterations, hence with $\boldsymbol{\epsilon} = 0.025 \cdot \mathbf{1}$. The same approximation was computed 100 times and we found it fairly stable (we obtained a standard deviation of the 100 replicates smaller than 2). We report the average estimate, which can be seen as a single estimate obtained from a concatenated sample. The other two Bayes factors do not involve about equality constraints, but only inequality constraints. For the case of PQD (i.e., $\hat{B}_{31}$), we did not need importance sampling because sampling directly from the prior and posterior gives a large number of draws satisfying the constraint.

The hypothesis of independence must be definitely rejected, whereas that of positive association may be accepted. In particular, the model incorporating TP_2, formulated by requiring that all the local log-odds ratios are non-negative, has to be preferred to that incorporating PQD, which is formulated through global log-odds ratios. This means that the data conform to the strongest notion of positive association among those considered by Douglas et al. (1990). Hence, we can state that sons coming from a better family have a higher chance of success also conditional on remaining within any given subset of neighboring classes. On the other hand, the hypothesis of uniform association has to be rejected since, comparing the model incorporating this constraint in addition to TP_2 (M_5) with model M_4, we obtained $\log(\hat{B}_{54}) = -28.01$.

In order to perform some sensitivity analysis, we also calculated the Bayes factors in the table above under other three different Dirichlet priors, obtaining the following results:

	$\log(\hat{B}_{21})$	$\log(\hat{B}_{31})$	$\log(\hat{B}_{41})$
$D(0.5 \cdot \mathbf{1})$	−34.49	4.26	5.04
$D(2 \cdot \mathbf{1})$	−34.22	4.36	5.26
$D(5 \cdot \mathbf{1})$	−32.18	4.39	6.04

It can be seen that there is only a slight sensitivity to prior assumptions for these data. By varying the prior parameters, we do not reach different conclusions with respect to model choice. We obtained similar results, not reported here, for the other Bayes factors computed in this section and in the next two sections.

Moving back to the data, we also considered some constraints on the marginal distributions of the response variables. In particular, we considered model M_6, formulated by incorporating in M_4 the constraint that the marginal distributions

Table 3
Alzheimer's disease (A_1) and cognitive impairment (A_2) for a sample of 513 elderly people, stratified by age (B: less than 75, more than 75).

A_1	A_2 (<75)				A_2 (≥75)			
	IV	III	II	I	IV	III	II	I
V	2	1	1	0	14	24	2	0
IV	1	12	10	1	19	48	25	0
III	0	8	27	5	1	25	63	4
II	0	0	20	4	0	0	35	7
I	0	0	0	85	0	0	0	69

of A_1 and A_2 are equal, and model M_7, by incorporating in M_4 the constraint that every local logit of A_2 is greater than the corresponding local logit of A_1; this in turn implies that the marginal distribution of A_2 is stochastically greater than that of A_1. The Bayes factors of these two models with respect to M_4 are:

$\log(\hat{B}_{64})$	$\log(\hat{B}_{74})$
−0.78	2.22

Model M_7 seems to be supported by the data. This means that we can observe not only *pure mobility*, that is positive association between family's origin and the son's status, but also *structural mobility*, which instead refers to how far apart the two marginal distributions are and is essentially related to socioeconomic growth.

4.2. Classification of elderly people by Alzheimer's disease and cognitive impairment

The second dataset we analyzed (see Table 3) is referred to a sample of elderly people cross-classified by Alzheimer's disease (A_1) and cognitive impairment (A_2), stratified by age (B); the data are taken from Agresti (2002, Chapter 7). The levels of A_1 are: (IV) highly probable; (III) probable; (II) possible; (I) unaffected; the levels of A_2 are: (V) severe; (IV) moderate; (III) mild; (II) borderline; (I) unaffected.

As the categories of both response variables are in reverse order, we based our analysis on reverse continuation logits. In this setting, we compared the saturated model (M_1) with the model of conditional independence (M_2) and the saturated model incorporating positive association in each stratum (M_3). The estimated Bayes factors are:

$\log(\hat{B}_{21})$	$\log(\hat{B}_{31})$
−6.31	4.76

The hypothesis of conditional independence is not supported by the data, whereas that of positive association in each stratum is strongly supported. This means that, also conditionally on the age, worst diagnoses of Alzheimer's disease are associated with most severe cognitive impairment.

We tried to test further hypotheses on the association between the two response variables. In particular, we considered the following constraints: (i) the level of the association is the same in each stratum; (ii) the association is stronger in the first stratum; (iii) the association is stronger in the second stratum. The models obtained by incorporating these hypotheses in M_3 are denoted by M_4, M_5, and M_6, respectively. The estimated Bayes factors for these models with respect to M_3 are:

$\log(\hat{B}_{43})$	$\log(\hat{B}_{53})$	$\log(\hat{B}_{63})$
−4.26	−22.40	8.12

A certain amount of evidence in favor of model M_6 is noted. Using this as reference model, further constraints on the marginal distributions can be added, such as: the marginal distribution of A_1 increases, namely the reverse continuation logits decreases, with age (M_7); the marginal distribution of A_2 increases with age (M_8); both marginal distributions increase with age (M_9). We have the following results:

$\log(\hat{B}_{76})$	$\log(\hat{B}_{86})$	$\log(\hat{B}_{96})$
6.37	17.17	22.14

Models M_7, M_8, and M_9 seem to be all compatible with the data, therefore we chose the one with the highest Bayes factor as the most plausible one, which is model M_9. This implies that, as age increases, individuals are more likely to have a serious level of cognitive impairment and to be diagnosed the Alzheimer's disease with a higher degree of confidence. Therefore, age does not only affect the association between the two response variables, that is stronger for elder people, but also shows a direct effect on their marginal distributions.

Table 4
Response to treatment over time (A_1, A_2, A_3, A_4) for a sample of 72 subjects, stratified by the type of treatment (B).

			Treatment									Placebo								
		A_3	I			II			III			I			II			III		
A_1	A_2	A_4	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
	I		0	1	0	0	0	0	0	0	0	0	6	1	0	2	0	0	0	0
I	II		0	2	0	0	3	2	0	0	1	0	3	1	0	6	2	0	0	0
	III		0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0
	I		0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
II	II		0	0	1	0	2	4	1	1	0	0	1	0	0	1	2	0	3	3
	III		0	0	0	0	1	3	0	0	5	0	0	0	0	1	0	0	0	1
	I		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
III	II		0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0
	III		0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0

4.3. Clinical trial for skin disorder

The dataset in Table 4, already analyzed by Glonek and McCullagh (1995) and Koch et al. (1991), refers to a clinical trial which, for confidentiality, was fictitiously described as pertaining to the treatment of a skin disorder. The 72 subjects in the study were divided into two groups, the first one receiving the treatment and the second one receiving the placebo. An ordinal response variable, with levels poor/fair (I), good (II), and excellent (III), was recorded for each subject on four different occasions: 3 days (A_1), 7 days (A_2), 10 days (A_3), and 14 days (A_4) after the treatment. Given the nature of the response variables, it is natural to use global logits.

The data are very sparse, as 128 of the 162 cells are empty. Therefore, following Glonek and McCullagh (1995), the largest model we considered is a reduced model in which all the interactions of order higher than two are set equal to zero. Such a model, that we indicate by M_2, does not constrain the association between every pair of response variables to be the same in the two strata and, in this respect, is less restrictive than the largest model considered by Glonek and McCullagh (1995). Note that M_1, the saturated model, is still used as a reference model to calculate the Bayes factors but is not included in the set of models of interest.

We first compared model M_2 with the model that Glonek and McCullagh (1995) chose as final model (M_3). The latter is based on the following constraints: (i) there is uniform association within each stratum and between the strata; (ii) there is a constant shift between marginal logits over time and between strata. Comparing this model, with model M_2, we obtained $\log(\hat{B}_{32}) = 0.19$. There is a very mild evidence in favor of M_3. Finally, using M_3 as reference model, we considered model M_4 obtained from M_3 by incorporating PQD and the constraint that the marginal distribution of each response variable is stochastically greater for the second stratum than for the first one. These hypotheses seem to be supported by the data, as we have $\log(\hat{B}_{43}) = 2.38$.

5. Discussion

In this paper, we propose a unifying framework for selecting marginal models (McCullagh and Nelder, 1989; Bergsma et al., 2009) for contingency tables which are specified through about equality and inequality constraints. The approach is based on a simple specification of prior distributions, for deriving coherent inference, and on the use of the Bayes factor.

Our proposed strategy involves only evaluation of the probability mass given to the constrained parameter space under the prior and under the posterior. In certain cases this is straightforward. In other cases, importance sampling is needed to overcome a rare event problem.

We have proposed using importance sampling with Dirichlet proposal distribution, and a strategy for obtaining a possibly good proposal by centering it around the maximum likelihood estimate. We underline that there is nothing making the Dirichlet distribution a good proposal in general. A great advantage of the Dirichlet distribution is that very efficient algorithms allow to quickly generate an extremely large number of independent draws from it. In our implementation, we have sometimes generated up to a million draws in few seconds, obtaining highly precise importance sampling estimates. The Dirichlet distribution can though be a good proposal only after careful tuning, and our maximum likelihood strategy does not necessarily give a satisfactory solution. We recommend checking the Effective Sample Size (Liu, 2001) after performing importance sampling and improving the proposal distribution if needed. A limitation of this procedure is that it is not automatic, and can be computationally intensive when repeated many times. Nevertheless, we suggest using the proposed approach, instead of the maximum likelihood approach in the following situations. First of all when comparing models parametrized through different types of logit, which would be otherwise cumbersome using the likelihood ratio test. Secondly, whenever dealing with frequencies or tables moderately large. The distribution of the likelihood ratio test depends, in fact, on the nuisance parameters represented by the marginal distributions. The conditional method used to deal with this problem under the maximum likelihood approach results in the multivariate generalized hypergeometric distribution (Bartolucci et al., 2001) which is intractable for moderately large tables and requires a computationally demanding Monte Carlo maximum likelihood approach for parameter estimation (Bartolucci and Scaccia, 2004). The approach we propose in

this paper, being based on the Bayes factor, is not affected by the presence of nuisance parameters and, in these cases, it represents a viable alternative to the maximum likelihood approach.

Appendix A. Transformation from π to η and its inversion

The matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ in (5) may be obtained as follows. $\boldsymbol{C}$ is a block diagonal matrix with blocks of columns $\boldsymbol{C}_{\boldsymbol{z}}$, ordered as the subvectors $\boldsymbol{\eta}_{\boldsymbol{z}}$ in $\boldsymbol{\eta}$, given by

$$\boldsymbol{C}_{\boldsymbol{z}} = \bigotimes_{i=1}^{q} \boldsymbol{C}_i,$$

where $\boldsymbol{C}_i = 1$ if $z_i = 1$ and $\boldsymbol{C}_i = \begin{pmatrix} -\boldsymbol{I}_{m_i-1} & \boldsymbol{I}_{m_i-1} \end{pmatrix}$ otherwise. Similarly, $\boldsymbol{M}$ has blocks of rows $\boldsymbol{M}_{\boldsymbol{z}}$ given by

$$\boldsymbol{M}_{\boldsymbol{z}} = \bigotimes_{i=1}^{q} \boldsymbol{M}_i,$$

where $\boldsymbol{M}_i = \boldsymbol{1}'$ if $z_i = 0$; otherwise, we have

$$\boldsymbol{M}_i = \begin{cases} \begin{pmatrix} \boldsymbol{I}_{m_i-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m_i-1} \end{pmatrix} & \text{if logits of type } l \text{ are used for the } i\text{-th variable,} \\ \begin{pmatrix} \boldsymbol{T}_{m_i-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}_{m_i-1}{}' \end{pmatrix} & \text{if logits of type } g \text{ are used for the } i\text{-th variable,} \\ \begin{pmatrix} \boldsymbol{I}_{m_i-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}_{m_i-1}{}' \end{pmatrix} & \text{if logits of type } c \text{ are used for the } i\text{-th variable,} \\ \begin{pmatrix} \boldsymbol{T}_{m_i-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m_i-1} \end{pmatrix} & \text{if logits of type } r \text{ are used for the } i\text{-th variable,} \end{cases}$$

where $\boldsymbol{T}_h$ is an $h \times h$ lower triangular matrix of ones.

In order to invert the transformation at issue, and then to obtain $\boldsymbol{\pi}$ corresponding to a given $\boldsymbol{\eta}$, the most general method is based on a Newton algorithm. In order to take into account that $\boldsymbol{\pi}$ belongs to the simplex of order r, this algorithm updates, instead of $\boldsymbol{\pi}$, a log-linear representation of this vector of type (2), which is based on $\boldsymbol{H}' = \begin{pmatrix} -\boldsymbol{1} & \boldsymbol{I}_{r-1} \end{pmatrix}$, so that

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{0}' \\ \boldsymbol{I}_{r-1} \end{pmatrix}.$$

This parametrization is based on a vector of log-linear parameters denoted by $\boldsymbol{\lambda}$ on the basis of which we obtain $\boldsymbol{\pi}$ through the explicit expression in (3).

Let $\boldsymbol{\lambda}^{(t)}$ denote the vector of log-linear parameters at the t-th step of the Newton algorithm, let $\boldsymbol{\pi}^{(t)}$ denote the corresponding joint probability vector, and let $\boldsymbol{\eta}^{(t)}$ denote the corresponding vector of marginal parameters obtained by (6). Then, starting from an initial value $\boldsymbol{\lambda}^{(0)}$, this algorithm iteratively updates it, by performing at the t-step the following operation:

$$\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^{(t-1)} + \boldsymbol{R}^{(t-1)}(\boldsymbol{\eta} - \boldsymbol{\eta}^{(t-1)}),$$

where

$$\boldsymbol{R}^{(t)} = [\boldsymbol{C}(\boldsymbol{M}\boldsymbol{\pi}^{(t)})^{-1}\boldsymbol{M}\mathrm{diag}(\boldsymbol{\pi}^{(t)})\boldsymbol{G}]^{-1}.$$

These steps are repeated until convergence, that is until the norm of $\boldsymbol{\pi}^{(t)} - \boldsymbol{\pi}^{(t-1)}$ becomes smaller than a certain threshold, such as 10^{-6}.

Appendix B. Computing Bayes factors with about equality constraints

First of all, we recall that about equality constraints are specified as $|\boldsymbol{E}\boldsymbol{\eta}| \leq \boldsymbol{\epsilon}$, for a small $\boldsymbol{\epsilon} > 0$. If $\boldsymbol{\epsilon}$ is too large, the corresponding Bayes factor is far from the Bayes factor which would be obtained with precise equality constraints. If $\boldsymbol{\epsilon}$ is too small, estimation of the proportion of the parameter space in agreement with the constraints, under the encompassing prior and the encompassing posterior distribution, may be inefficient.

In order to fix a suitable value for $\boldsymbol{\epsilon}$, we adapt the iterative procedure of Klugkist et al. (2010). Suppose we want to estimate B_{k1} for the constrained model M_k versus the encompassing model, where the constrained model is specified by $|\boldsymbol{E}\boldsymbol{\eta}| \leq \boldsymbol{\epsilon}$ and, possibly, $\boldsymbol{U}\boldsymbol{\eta} \geq \boldsymbol{0}$. Our procedure is based on the following steps:

1. choose a small value $\boldsymbol{\epsilon}_1$ and define $M_{k,1}$ as the model M_k in which $\boldsymbol{\epsilon}$ is put equal to $\boldsymbol{\epsilon}_1$;
2. estimate $\hat{B}_{(k,1)1} = \hat{\alpha}_{k,1}/\hat{\beta}_{k,1}$, where $1/\hat{\alpha}_{k,1}$ and $1/\hat{\beta}_{k,1}$ are the proportions of the samples from the encompassing prior and posterior distributions in agreement with the constraints used to formulate $M_{k,1}$, respectively;

3. define $\boldsymbol{\epsilon}_2 = b\boldsymbol{\epsilon}_1$, with $0 < b < 1$, and $M_{k,2}$ as the model M_k in which $\boldsymbol{\epsilon}$ is put equal to $\boldsymbol{\epsilon}_2$;
4. estimate $\hat{B}_{(k,2)(k,1)} = (\hat{\alpha}_{k,2}/\hat{\beta}_{k,2})/(\hat{\alpha}_{k,1}/\hat{\beta}_{k,1})$, where $1/\hat{\alpha}_{k,2}$ and $1/\hat{\beta}_{k,2}$ are the proportions of the samples from the encompassing prior and posterior in agreement with the constraints used to formulate $M_{k,2}$, respectively.

Repeat steps 3 and 4, with each $\boldsymbol{\epsilon}_{j+1} = b\boldsymbol{\epsilon}_j$, until the condition $\hat{B}_{(k,j+1)(k,j)} \approx 1$ is satisfied. Then the required Bayes factor estimate $\hat{B}_{k1}$ can be calculated by multiplication:

$$\hat{B}_{k1} = \hat{B}_{(k,1)1} \times \hat{B}_{(k,2)(k,1)} \times \cdots \times \hat{B}_{(k,j)(k,j-1)} = \hat{\alpha}_{k,j}/\hat{\beta}_{k,j}. \tag{B.1}$$

In the limit (i.e., when $\boldsymbol{\epsilon}_j \to \mathbf{0}$), this method yields the estimate of the Bayes factor for model M_k with exact equality constraints versus the encompassing model.

Note that, in the procedure above, the problem of getting inefficient estimates for the proportion of encompassing prior and posterior in agreement with the constraints is solved by using the importance sampling approach described in Section 3.3. Thus, only two different samples, one drawn from the importance density for the prior and the other one from the importance density for the posterior, are required to compute all the Bayes factor estimates in (B.1).

References

Agresti, A., 2002. Categorical Data Analysis, second ed. John Wiley & Sons, New York.
Albert, J.H., 1996. Bayesian selection of log-linear models. Canadian Journal of Statistics 24, 327–347.
Albert, J.H., 1997. Bayesian testing and estimation of association in a two-way contingency table. Journal of the American Statistical Association 92, 685–693.
Bartolucci, F., Colombi, R., Forcina, A., 2007. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. Statistica Sinica 17, 691–711.
Bartolucci, F., Forcina, A., Dardanoni, V., 2001. Positive quadrant dependence and marginal modeling in two-way tables with ordered margins. Journal of the American Statistical Association 96, 1497–1505.
Bartolucci, F., Scaccia, L., 2004. Testing for positive association in contingency tables with fixed margins. Computational Statistics and Data Analysis 47, 195–210.
Berger, J.O., Delampady, M., 1987. Testing precise hypotheses. Statistical Science 2, 317–352.
Berger, J.O., Sellke, T., 1987. Testing a point null hypothesis: the irreconcilability of p values and evidence. Journal of the American Statistical Association 82, 112–122.
Bergsma, W., Croon, M., Hagenaars, J.A., 2009. Marginal Models: For Dependent, Clustered, and Longitudinal Categorical Data. Springer, New York.
Bergsma, W.P., Rudas, T., 2002. Marginal models for categorical data. The Annals of Statistics 30, 140–159.
Bucklew, J., 2004. Introduction to Rare Event Simulation. Springer, New York.
Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.
Christensen, R., 1997. Log-Linear Models and Logistic Regression, second ed. Springer, New York.
Colombi, R., Forcina, A., 2001. Marginal regression models for the analysis of positive association of ordinal response variables. Biometrika 88, 1007–1019.
Dardanoni, V., Forcina, A., 1998. A unified approach to likelihood inference on stochastic orderings in a nonparametric context. Journal of the American Statistical Association 93, 1112–1123.
Dawid, A.P., Lauritzen, S.L., 2001. Compatible prior distributions. In: George, E.I. (Ed.), Bayesian Methods with Applications to Science, Policy and Official Statistics. Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis. Eurostat, Luxembourg, pp. 109–118.
Dellaportas, P., Forster, J.J., 1999. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86, 615–633.
Douglas, R., Fienberg, S.E., Lee, M.-L.T., Sampson, A.R., Whitaker, L.R., (1990). Positive dependence concepts for ordinal contingency tables. In: H.W. Block, A.R. Sampson and T.H. Savits (Eds.), Topics in Statistical Dependence, Hayward, CA, pp. 189–202.
Gamerman, D., Lopes, H.F., 2006. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, second ed. Chapman & Hall, CRC, Boca Raton, FL.
Glonek, G.F.V., 1996. A class of regression models for multivariate categorical responses. Biometrika 83, 15–28.
Glonek, G.F.V., McCullagh, P., 1995. Multivariate logistic models. Journal of the Royal Statistical Society: Series B 57, 533–546.
Goodman, L., 1991. Measures, models, and graphical displays in the analysis of cross-classified data. Journal of the American Statistical Association 86, 1085–1111.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Hoijtink, H., Klugkist, I., Boelen, P., 2008. Bayesian Evaluation of Informative Hypotheses. Springer, New York.
Jefferys, W.H., Berger, J.O., 1992. Ockham's razor and Bayesian analysis. American Scientist 80, 64–72.
Jeffreys, H., 1935. Some tests of significance, treated by the theory of probability. Mathematical Proceedings of the Cambridge Philosophical Society 31, 203–222.
Jeffreys, H., 1961. Theory of Probability, third ed. Oxford University Press.
Karlin, S., 1968. Total Positivity. Stanford University Press.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Klugkist, I., Hoijtink, H., 2007. The Bayes factor for inequality and about equality constrained models. Computational Statistics and Data Analysis 51, 6367–6379.
Klugkist, I., Kato, B., Hoijtink, H., 2005a. Bayesian model selection using encompassing priors. Statistica Neerlandica 59, 57–69.
Klugkist, I., Laudy, O., Hoijtink, H., 2005b. Inequality constrained analysis of variance: a Bayesian approach. Psychological Methods 10, 477–493.
Klugkist, I., Laudy, O., Hoijtink, H., 2010. Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. Psychologial Methods 15, 281–299.
Koch, G., Singer, J., Stokes, M., Carr, G., Cohen, S., Forthofer, R., 1991. Some aspects of weighted least squares analysis for longitudinal categorical data. In: Dwyer, J.H., Feinlib, M., Lippert, P., Hoffmeister, H. (Eds.), Statistical Models for Longitudinal Studies of Health. Oxford University Press, pp. 215–258.
Lang, B.J., 1996. Maximum likelihood methods for a generalized class of log-linear models. The Annals of Statistics 24, 726–752.
Lehmann, E.L., 1966. Some concepts of dependence. The Annals of Mathematical Statistics 37, 1137–1153.
Liu, J.S., 2001. Monte Carlo Strategies in Scientific Computing. Springer, New York.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, London.
Plackett, R.L., 1965. A class of bivariate distributions. Journal of the American Statistical Association 60, 516–522.
Spiegelhalter, D.J., Smith, A.F.M., 1982. Bayes factors for linear and log-linear models with vague prior information. Journal of the Royal Statistical Society: Series B 44, 377–387.
Tuyl, F., Gerlach, R., Mengersen, K., 2009. Posterior predictive arguments in favor of the Bayes–Laplace prior as the consensus prior for binomial and multinomial parameters. Bayesian Analysis 4, 151–158.
Wetzels, R., Grasman, R.P.P.P., Wagenmakers, E.-J., 2010. An encompassing prior generalization of the Savage–Dickey density ratio. Computational Statistics and Data Analysis 54, 2094–2102.