Bayesian Inference for the RC(*m*) Association Model

Maria KATERI, Anna NICOLAOU, and Ioannis NTZOUFRAS

Describing the structure in a two-way contingency table in terms of an RC(m) association model, we are concerned with the computation of posterior distributions of the model parameters using prior distributions which take into account the nonlinear restrictions of the model. We are further involved with the determination of the order of association m, based on Bayesian arguments. Using projection methods, a prior distribution over the parameters of the simpler RC(m) model is induced from a prior of the parameters of the saturated model. The fit of the assumed RC(m) model is evaluated using the posterior distribution of its distance from the full model. Our methods are illustrated with a popular dataset.

Key Words: Contingency tables; Kullback–Leibler projection; MCMC methods.

1. INTRODUCTION

Let $\mathbf{\Pi} = (\pi_{ij})$ be the $I \times J$ probability table corresponding to a frequency table $\boldsymbol{y} = (y_{ij})$ with $n = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$, that is produced by the cross-classification of two categorical variables with I and J categories, respectively. The underlying sampling scheme can be either multinomial with expected cell frequencies $\xi_{ij} = n\pi_{ij}$ or independent Poisson for each cell of the table with means ξ_{ij}. The saturated log-linear model is expressed as

$$\eta_{ij} = \log(\xi_{ij}) = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \lambda_{ij}^{(12)} \quad i = 1, \ldots, I,\ j = 1, \ldots, J\ , \tag{1.1}$$

where the row and column main effects, $\lambda_i^{(1)}$ and $\lambda_j^{(2)}$ sum over i and j, respectively, to zero. The zero-sum constraints (over i and j) hold also for the interaction parameters $\lambda_{ij}^{(12)}$. The interaction terms matrix is of full rank $M^* = \min(I, J) - 1$. Hence, we may consider

Maria Kateri is Assistant Professor, Department of Statistics and Insurance Science, University of Piraeus, Greece (E-mail: mkateri@unipi.gr). Anna Nicolaou is Assistant Professor, Department of Business Administration, University of Macedonia, Thessaloniki, Greece (E-mail: nicolaou@uom.gr). Ioannis Ntzoufras is Assistant Professor, Department of Statistics, Athens University of Economics and Business, Athens, Greece (E-mail: ntzoufras@aueb.gr).

Journal of Computational and Graphical Statistics, Volume 14, Number 1, Pages 116–138
DOI: 10.1198/106186005X24944

its generalized singular value decomposition

$$\lambda_{ij}^{(12)} = \sum_{k=1}^{M^*} \phi_k \mu_{ik} \nu_{jk},$$

where $\phi_1 \geqslant \cdots \geqslant \phi_{M^*} > 0$ and the row and column scores μ_{ik}, ν_{jk} satisfy the constraints

$$\sum_{i=1}^{I} w_{1i} \mu_{ik} = \sum_{j=1}^{J} w_{2j} \nu_{jk} = 0\,, \quad k = 1, \ldots, M^*\,, \tag{1.2}$$

$$\sum_{i=1}^{I} w_{1i} \mu_{ik} \mu_{i\ell} = \sum_{j=1}^{J} w_{2j} \nu_{jk} \nu_{j\ell} = \delta_{k\ell}\,, \quad k, \ell = 1, \ldots, M^*\,.$$

w_{1i}, w_{2j} are some positive weights and $\delta_{k\ell}$ denotes Kronecker's δ. Common values of the weights are $w_{1i} = w_{2j} = 1$, for all i, j or $w_{1i} = \pi_{i\cdot}$ and $w_{2j} = \pi_{\cdot j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. This decomposition leads to a reparameterization of (1.1) called RC(M^*) association model. If we retain only the first m eigenvalues, assuming that $\phi_{m+1} = \cdots = \phi_{M^*} = 0$, we obtain the association model RC(m) of mth order (Goodman 1985)

$$\eta_{ij} = \log(\xi_{ij}) = \lambda + \lambda_i^{(1)} + \lambda_j^{(2)} + \sum_{k=1}^{m} \phi_k \mu_{ik} \nu_{jk}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J\,. \tag{1.3}$$

The representation of the log-odds ratio for any 2×2 subtable formed from rows i and i' and columns j and j' as a sum of m components

$$\log\left(\frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}}\right) = \sum_{k=1}^{m} \phi_k (\mu_{ik} - \mu_{i'k})(\nu_{jk} - \nu_{j'k})$$

offers an insight into the meaning of these parameters. Another interpretation of the quantity ϕ_m is provided by expressing it as a contrast of the parameters η_{ij}'s (Goodman 1991)

$$\phi_m = \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{im} \nu_{jm} \eta_{ij}.$$

For $m = 1$, (1.3) reduces to the multiplicative row-column association model RC. In the case of ordered contingency tables, if we assign to the rows and columns of the table scores that reflect their ordinality, we obtain the uniform association model, with one parameter additional to the independence model. Bayesian methods that incorporate the category orderings have been proposed by Chuang (1982); Agresti and Chuang (1989); Evans, Gilula, and Guttman (1993); and Albert (1997). Agresti and Chuang (1989) considered a prior distribution for the cell proportions or for the parameters of the saturated log linear model and chose the components of the prior mean to satisfy a simple model for ordinal data such as the uniform association model. Evans et al. (1993) set a prior distribution for the parameters of the saturated log linear model and then they estimated the posterior distribution of

the parameters of the RC(1) model using the values that minimize the Euclidean squared distance between the interaction terms of the two models. As a referee pointed out, this approach can be extended in RC(m) models. However, their approach can be thought of as only an approximation of the correct posterior distribution. Furthermore, such an approach does not allow for the implementation of the full Bayesian inference including evaluation of the posterior model weights and implementation of Bayesian model averaging.

The contribution of this article is two-fold. First, we present a Markov chain Monte Carlo (MCMC) algorithm that is used to compute the posterior distributions of the parameters of the RC(m) model. Second, we are concerned with the assessment of the fit of such models. The methodology we introduce in this article is important because it can be considered as the first step for full Bayesian analysis in RC(m) models. Further work on this aspect may include incorporation of prior information, estimation of posterior model probabilities, and Bayesian model averaging. In Section 2, we propose vague prior distributions for the parameters of the RC(m) model that take into account the nonlinear restrictions of the model and we describe in detail the MCMC algorithm. We proceed in Section 3 using projection methods to evaluate the fit of the interaction components and determine the number of components that are required to produce a reasonable approximation of the data. We propose rules to assess the fit of the RC(m) model based on the posterior distribution of its Kullback–Leibler distance from the saturated RC(M^*). We use a popular dataset to illustrate our results in Section 4. We conclude with a discussion on the issues of incorporation and elicitation of prior information as well as the extension of the MCMC methodology over parameter spaces of different dimension.

2. BAYESIAN ESTIMATION OF THE RC(m) MODEL THROUGH MARKOV CHAIN MONTE CARLO

Without loss of generality, we consider the Poisson sampling. The likelihood function of the corresponding RC(m) model is

$$f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \boldsymbol{\Phi}_m) = \exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J} e^{\eta_{ij}} + \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\eta_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{J} \log(y_{ij}!)\right),$$

where η_{ij} is given by (1.3), $\boldsymbol{\Lambda}^{(1)} = \left(\lambda_i^{(1)}\right)$ and $\boldsymbol{\Lambda}^{(2)} = \left(\lambda_j^{(2)}\right)$ are the $I \times 1$ and $J \times 1$ vectors of the row and column main effects, $\mathbf{M}_m = (\mu_{ik})$, $\mathbf{N}_m = (\nu_{jk})$ are the $I \times m$ and $J \times m$ matrices of the row and columns scores, and $\boldsymbol{\Phi}_m = \text{diag}(\phi_1, \ldots, \phi_m)$. For the parameters of the main effects we use sum-to-zero constraints and fairly "vague" normal prior distributions. Therefore we set

$$\lambda_1^{(1)} = -\sum_{i=2}^{I} \lambda_i^{(1)}, \quad \lambda_1^{(2)} = -\sum_{j=2}^{J} \lambda_j^{(2)}$$

and we assume that $\lambda \sim \text{Normal}(0, \sigma_0^2)$, $\lambda_i^{(1)} \sim \text{Normal}(0, \sigma_1^2)$, $\lambda_j^{(2)} \sim \text{Normal}(0, \sigma_2^2)$ for $i = 2, \ldots, I$ and $j = 2, \ldots, J$. In our examples we select $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 1{,}000$.

For the distribution of ϕ_1 we use the log-normal with prior mean $\mu_{\phi_1} = 0$ and variance $\sigma_{\phi_1}^2$ of the $\log \phi_1$. In our example we set $\sigma_{\phi_1}^2 = 1{,}000$. All the other conditional distributions $f(\phi_k|\phi_{k-1})$, $k = 2, \ldots, m$, are defined as uniform distributions on the intervals $(0, \phi_{k-1})$. Hence, the joint prior distribution is given by

$$\begin{aligned} f(\phi_1, \ldots, \phi_m) &= f(\phi_m|\phi_1, \ldots, \phi_{m-1}) f(\phi_{m-1}|\phi_1, \ldots, \phi_{m-2}), \ldots f(\phi_2|\phi_1) f(\phi_1) \\ &= f(\phi_1) \prod_{k=2}^{m} \frac{1}{\phi_{k-1}} \mathcal{I}(\, 0 < \phi_k < \phi_{k-1}\,), \end{aligned}$$

where $\mathcal{I}(X)$ is equal to one if X is true and zero otherwise.

For the row and column scores $\mathbf{M}_m$ and $\mathbf{N}_m$, respectively, we use the uniform prior suggested by Viele and Srinivasan (2000) in the context of two-way analysis of variance models with multiplicative interaction terms. The columns $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{Ik})^T$ $k = 1, \ldots, m$ of the matrix $\mathbf{M}_m$ (similarly for $\mathbf{N}_m$) are generated successively to satisfy the constraints (1.2). Each score vector has unit length and therefore it can be visualized as a point on the surface of a hypersphere of unit radius; its elements sum to zero implying that it is orthogonal to the unit vector and they are also orthogonal to each other. Hence, we can regard the first column vector $\boldsymbol{\mu}_1$ as a point on the surface of a sphere that is the intersection of an I dimensional hypersphere S_I of unit radius with a hyperplane in R^I orthogonal to the unit vector. The additional score vectors need to be also orthogonal to all the previous score vectors and so they reside in lower dimensional spheres. We assume that each of the $\boldsymbol{\mu}_k$ $k = 1, \ldots, m$ is uniformly distributed on the surface of the sphere where it takes its values

$$f\left(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m\right) = \left\{ \prod_{k=1}^{m} (2\pi)^{-\frac{1}{2}(I-k)} \Gamma\left(\frac{I-k}{2}\right) \right\} \mathcal{I}\left(\mathbf{M}_m^T \mathbf{M}_m = \boldsymbol{I}_I\right),$$

where $\boldsymbol{I}_I$ is the $I \times I$ unit matrix. To generate values for $\boldsymbol{\mu}_1$ we first introduce a latent variable $\boldsymbol{\tau}_1$ that is uniformly distributed on the surface of a sphere that is the intersection of S_I with a hyperplane orthogonal to one of the basis vectors of R^I, say the first. So its first element is $\tau_{11} = 0$ and the rest are generated as $\tau_{l1} = u_l / \sqrt{\sum_{l=2}^{I} u_l^2}$ where $u_2, \ldots, u_I$ are independent normal variables with zero mean and variance one. Then, we transform to $\boldsymbol{\mu}_1 = \boldsymbol{D}\boldsymbol{\tau}_1$ multiplying by a matrix whose columns define an orthonormal basis of R^I and its first column is the unit vector. The additional columns of $\mathbf{M}_m$ are generated similarly orthogonal to the unit vector and also to all the previously generated score vectors.

Construction of an MCMC algorithm for the RC(m) model is not a straightforward task. This is mainly due to the multiplicative nature of the interaction parameters and the complicated constraints imposed on them. We describe the MCMC algorithm in as much detail as possible in order to initiate further improvements or alternative approaches on the topic. Our MCMC algorithm is summarized by the following steps.

1. Sample λ, $\lambda_i^{(1)}$, $i = 2, \ldots, I$ and $\lambda_j^{(2)}$, $j = 2, \ldots, J$ from the corresponding conditional posterior distributions which are proportional to

$$f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \boldsymbol{\Phi}_m) \times f(\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}).$$

2. For $k = 1, \ldots, m$ sample ϕ_k from

$$f(\phi_k|\boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, [\boldsymbol{\Phi}_m]_{\setminus k}, \boldsymbol{y}) \propto f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \boldsymbol{\Phi}_m) \times f(\boldsymbol{\Phi}_m),$$

where $[\boldsymbol{\Phi}_m]_{\setminus k}$ is a vector containing all diagonal elements of $\boldsymbol{\Phi}_m$ except ϕ_{kk}. The following Metropolis-Hastings step was used:

(a) If $k = 1$, then:

i. Propose $\phi_1' = \phi_2 + (\phi_1 - \phi_2)\exp(cu)$, where $u \sim N(0, 1)$ and c is a tuning parameter. This proposal induces a random walk for the parameter

$$\theta_1 = \log(\phi_1 - \phi_2). \tag{2.1}$$

ii. Accept the proposed value with probability $\alpha = \min(1, A)$, with A given by

$$A = \frac{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1', \phi_2, \ldots, \phi_m)}{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \phi_2, \ldots, \phi_m)} \times \frac{f(\phi_1')\phi_1}{f(\phi_1)\phi_1'} \times \frac{\phi_1' - \phi_2}{\phi_1 - \phi_2}.$$

(b) For $k = 2, \ldots, m-1$, then:

i. Propose

$$\phi_k' = \frac{\phi_{k+1} + \phi_{k-1}\frac{\phi_k - \phi_{k+1}}{\phi_{k-1} - \phi_k}\exp(cu)}{1 + \frac{\phi_k - \phi_{k+1}}{\phi_{k-1} - \phi_k}\exp(cu)},$$

where $u \sim N(0, 1)$ and c is a tuning parameter. This proposal induces a random walk for parameters

$$\theta_k = \log\left(\frac{\phi_k - \phi_{k+1}}{\phi_{k-1} - \phi_k}\right). \tag{2.2}$$

ii. Accept the proposed value with probability $\alpha = \min(1, A)$ with A given by

$$A = \frac{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_k', \ldots, \phi_m)}{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_k, \ldots, \phi_m)} \times \frac{\phi_k}{\phi_k'} \times \frac{(\phi_k' - \phi_{k+1})(\phi_{k-1} - \phi_k')}{(\phi_k - \phi_{k+1})(\phi_{k-1} - \phi_k)}.$$

(c) For $k = m$:

i. Propose

$$\phi'_m = \frac{\phi_{m-1}\frac{\phi_m}{\phi_{m-1}-\phi_m}\exp(cu)}{1+\frac{\phi_m}{\phi_{m-1}-\phi_m}\exp(cu)},$$

where $u \sim N(0,1)$ and c is a tuning parameter. This proposal induces a random walk for the parameter

$$\theta_m = \log\left(\frac{\phi_m}{\phi_{m-1}-\phi_m}\right). \tag{2.3}$$

ii. Accept the proposed value with probability $\alpha = \min(1, A)$ with A given by

$$A = \frac{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi'_m)}{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_m)} \times 1 \times \frac{\phi'_m(\phi_{m-1}-\phi'_m)}{\phi_m(\phi_{m-1}-\phi_m)}.$$

3. Update the row scores $\boldsymbol{M}_m$ using the following Metropolis step:

(a) For $k = 1, \ldots, m$ propose new row scores using the following steps:

i. For $\ell = 1, \ldots, k$ set $\tau'_{\ell k} = 0$.
For $\ell = k+1, \ldots, I$ sample. $\tau'_{\ell k} \sim N(\tau_{\ell k}, c^2_{\tau_{\ell k}})$.

ii. Rescale $\tau'_{\ell k}$ to add to unit: $\tau''_{\ell k} = \tau'_{\ell k}/\sqrt{\sum_{\ell=1}^{I}[\tau'_{\ell k}]^2}$.

iii. If $k = 1$, then set $\boldsymbol{B} = (\boldsymbol{1}_I, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_I)$ equal to $\boldsymbol{A} = (\boldsymbol{1}_I, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_I)$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are $I \times I$ matrices and $\boldsymbol{a}_j$'s are $I \times 1$ vectors with

$$a_{ij} = \begin{cases} 0 & \text{for} \quad i \le j-2 \\ I-i & \text{for} \quad i = j-1 \\ -1 & \text{for} \quad i \ge j \end{cases}$$

If $k \ge 2$, then for $j = 2, \ldots, k$ set $\boldsymbol{b}_j = \boldsymbol{\tau}''_{j-1}$ and $\boldsymbol{b}_1 = \boldsymbol{1}_I$,
for $j = k+1, \ldots, I$ set $\boldsymbol{b}_j = \boldsymbol{a}_{j-k+1}$.

iv. Implement Gram-Schmidt orthogonalization of the columns of $\boldsymbol{B}$ and produce matrix $\boldsymbol{C}$.

v. Normalize the columns of $\boldsymbol{C}$ and produce $\boldsymbol{D}$ with elements

$$d_{ij} = \frac{c_{ij}}{\sqrt{\sum_{\ell=1}^{I} c^2_{\ell j}}}.$$

vi. Set $\boldsymbol{\mu}'_k = \boldsymbol{D}\boldsymbol{\tau}''_k$, where $\boldsymbol{\tau}_{\ell k}$ are latent variables calculated in Step 3a(i).

(b) Accept the proposed move with probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}'_m, \boldsymbol{N}_m, \boldsymbol{\Phi}_m)}{f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \boldsymbol{\Phi}_m)}\right).$$

(c) If the proposed move is accepted, then set $\boldsymbol{M}_m = \boldsymbol{M}'_m$ and $\boldsymbol{\tau}_k = \boldsymbol{\tau}''_k$ for all $k = 1, \ldots, m$ else $\boldsymbol{M}_m$ and $\boldsymbol{\tau}_k$ remain unchanged.

4. Update the column scores using the same procedure as in Step 3.

In Step 1 we use a Metropolis random-walk algorithm. In Step 2 we employ multiplicative variations of it. Namely, we transform the parameters $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_m > 0$ using monotone, invertible transformations to new parameters $\theta_1, \ldots, \theta_m$ which lie in the whole real line (see Equations (2.1), (2.2), and (2.3)). Then, we generate values for the new parameters using the Metropolis algorithm and we obtain values of the original parameters by inverse transformation. More explicitly, we propose values for the transformed parameters $\theta_k' = \theta_k + cu$, where $u \sim N(0,1)$ and we accept the proposed move from θ_k to θ_k' with probability $\alpha = \min(1, A)$, where A is given by

$$\begin{aligned} A &= \frac{f(\theta_k'|\boldsymbol{y}, \lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_{k-1}, \phi_{k+1} \ldots, \phi_m)}{f(\theta_k|\boldsymbol{y}, \lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_{k-1}, \phi_{k+1} \ldots, \phi_m)} \\ &= \frac{f(\phi_k'|\boldsymbol{y}, \lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_m) \left|\frac{d\theta_k'}{d\phi_k'}\right|}{f(\phi_k|\boldsymbol{y}, \lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{M}_m, \boldsymbol{N}_m, \phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_m) \left|\frac{d\theta_k}{d\phi_k}\right|}. \end{aligned}$$

In all cases the variance of the proposal density c was tuned to achieve a 30–50% acceptance rate. Alternatively, for Steps 1 and 2, we may use the adaptive-rejection algorithm of Gilks and Wild (1992) for log-concave distributions.

For Steps 3 and 4 we use the approach of Viele and Srinivasan (2000). Note that they proposed multiple updates within each Gibbs step to improve the mixing of the chain. In our example we have updated 30 times the row and column scores in each iteration and this was proved sufficient. Furthermore, Viele and Srinivasan (2000) did not constrain the row and column scores in terms of signs arguing that the chain sticks with one set of signs. However, we did not observe the same behavior in our illustrative example. In order to avoid complicating the MCMC algorithm we also left the signs of the scores unconstrained within the MCMC algorithm and, in the output analysis, we transformed them suitably by setting $\mu_{1k} > 0$ and $\nu_{1k} > 0$ for $k = 1, 2, \ldots, M$. This was achieved by changing appropriately the sign of the rest of the μ_{ik}'s and ν_{ik}'s when necessary. Although it is generally undesirable to have nonidentifiable contrasts of parameters, as we do here, it is possible to help the stability of estimates (Spiegelhalter, Thomas, Best, and Gilks 1996, p. 53). The posterior densities of the score parameters are bimodal distributions with the modes corresponding to identical models. An alternative approach is to restrict the prior distribution only to the half-sphere by setting the first component of each $\boldsymbol{\nu}_k$ and $\boldsymbol{\mu}_k$ (for $k = 1, \ldots, m$) to be positive. Such an action though will considerably complicate Step 3 and slow down the MCMC algorithm. An easier but ad-hoc procedure is to transform the parameters after each iteration within the MCMC algorithm (Vines, Gilks, and Wild 1996, p. 342); that is, change the signs after each iteration and continue the algorithm with the transformed values.

We must further add that the shape of the posterior distribution is highly complicated due to the imposed constraints. Therefore, the chain needs to run long enough to explore the whole parameter space. The posterior mean or median may not be adequate descriptive measures of the posterior distribution because they do not satisfy the parameter restrictions and, moreover, the marginal distributions are highly skewed or, in some cases, possibly bimodal. Viele and Srinivasan (2000) proposed to consider the posterior means of the latent

parameters τ and transform them to row and column scores. We propose instead to monitor both, the posterior descriptive values and the density estimators, in order to have a better picture of the whole posterior distribution.

3. EVALUATION OF THE ORDER OF ASSOCIATION

In this section we deal with the problem of choosing and evaluating the fit of the order m of an RC(m) model. Generally, Bayesian inference concerning the order m should be based on the posterior model probability $f(m|\boldsymbol{y})$ or on the posterior model odds when we compare two models of different order. Methods to avoid the direct computation of the posterior model probabilities that often are not analytically tractable, are reviewed by Chipman, George, and McCulloch (2001) and Lopes (2002). In what follows, we facilitate an alternative and simpler approach using projection methods.

3.1 COMPUTING THE KULLBACK-LEIBLER PROJECTION

Following Goutis and Robert (1998) and McCulloch and Rossi (1993) we project the parameter vector $\boldsymbol{\theta}$ of the saturated RC(M^*) model to a vector of lower dimension $\boldsymbol{\theta}_m$ that parameterizes the distribution of the RC(m) model. The projection is defined by choosing $\boldsymbol{\theta}_m$ so as to minimize the Kullback–Leibler distance between the sampling distributions of the two models

$$\underline{\mathrm{KL}}_m = \min_{\boldsymbol{\theta}_m \in \boldsymbol{\Theta}_m} \mathrm{KL}(M^*, m) = \min_{\boldsymbol{\theta}_m \in \boldsymbol{\Theta}_m} \int f(\boldsymbol{y}|\boldsymbol{\theta}) \log \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta}_m)} dy. \tag{3.1}$$

For the simple RC(1) model, Evans et al. (1993) used the Euclidean projection. For any prior distribution on the parameters of the full model, the projection induces a prior over the restricted parameter space that generates a posterior distribution for $\boldsymbol{\theta}_m$. From this we deduce the posterior distribution of the minimum distance (3.1) based on which we judge whether or not the reduced model causes a small change in the goodness of fit of the observed data.

The Kullback–Leibler distance for the Poisson and the multinomial sampling differs by a constant term only. For an $I \times J$ contingency table, it can be written explicitly as

$$\mathrm{KL}(M^*, m) = \sum_{i=1}^{I} \sum_{j=1}^{J} \xi_{ij} \log \left(\frac{\xi_{ij}}{\xi_{ij}^m} \right) - \mathcal{I}_p \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\xi_{ij} - \xi_{ij}^m \right), \tag{3.2}$$

where ξ_{ij}^m are the expected cell frequencies under the RC(m) and ξ_{ij} are the frequencies of the saturated model. $\mathcal{I}_p$ is a binary index taking the values one or zero for the Poisson and the multinomial likelihood, respectively. The second term of (3.2) is zero for the multinomial case, because the total number of expected counts is considered constant, equal to the total sample size $n = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$. Under the Poisson sampling scheme, the minimum of this term with respect to $\boldsymbol{\theta}_m$ equals zero and thus both sampling schemes are treated simultaneously. The following lemma, being an extension of an associated result for generalized

linear models given by Goutis and Robert (1998), provides the calculation of the projected parameter vector $\boldsymbol{\theta}_m$.

Lemma 1. *Through the Kullback–Leibler projection method, the parameters of the RC(m) model are obtained from the true underlying model RC(M^*) as the solution of the system of equations*

$$\begin{aligned}
\xi_{i\cdot} &= \xi^m_{i\cdot}\,, \qquad i = 1,\ldots,I\,,\\
\xi_{\cdot j} &= \xi^m_{\cdot j}\,, \qquad j = 1,\ldots,J\,,\\
\sum_j \xi_{ij}\nu^m_{jk} &= \sum_j \xi^m_{ij}\nu^m_{jk}\,, \quad i = 1,\ldots,I\,,\ \ k = 1,\ldots,M\\
\sum_i \xi_{ij}\mu^m_{ik} &= \sum_i \xi^m_{ij}\mu^m_{ik}\,, \quad j = 1,\ldots,J\,,\ \ k = 1,\ldots,M\\
\sum_{i,j} \xi_{ij}\mu^m_{ik}\nu^m_{jk} &= \sum_{ij} \xi^m_{ij}\mu^m_{ik}\nu^m_{jk} \quad , \ \ k = 1,\ldots,M
\end{aligned} \tag{3.3}$$

which are the likelihood equations for RC(m) with the sampling frequencies (y_{ij}) replaced by the expected frequencies (ξ_{ij}).

Remark. Maximum likelihood estimation of the parameters of the RC(m) model can be achieved using the iterative algorithms of Becker (1990) and Haberman (1995). An immediate consequence of Lemma 1 is that $\boldsymbol{\theta}_m$ is computed by standard programs for the maximum likelihood fit of the RC(m) model applied on (π_{ij}) instead of the data $y_{ij}/(IJ)$.

The independence model RC(0), obtained by eliminating the sum in (1.3), is a standard reference model for association models. It will be illuminating to express $\underline{\mathrm{KL}}_m$ in terms of departures from the independence model.

Lemma 2. *The value $\boldsymbol{\theta}_m$ that achieves the minimum Kullback–Leibler distance KL(M^*, m) satisfies the relation*

$$KL(M^*, m) = KL(M^*, 0) - KL(m, 0), \tag{3.4}$$

when we use the marginal weights; where $m = 0$ stands for the independence model.

The above relation is a Pythagorean-type equality and is natural for the Kullback–Leibler information (Gokhale and Kullback 1978). An analogous equality is provided by Soofi (1992) for the conditional logit model. The distance $\mathrm{KL}(M^*, m)$ is positive and decreasing in m for $0 \leqslant m < M^*$; it is upper bounded by $\mathrm{KL}(M^*, m_{\text{con}}) = n\log(IJ)$, where m_{con} is the constant model with $\pi_{ij} = 1/(IJ)$ (Dupuis 1997). Thus, if $\mathrm{KL}(M^*, m) = 0$ the order of the association is m and $\mathrm{KL}(M^*, k) = 0$ for $k \geq m$.

3.2 CALIBRATING THE KULLBACK-LEIBLER DISTANCE

The process of generating a sample from the posterior distribution of $\underline{\mathrm{KL}}_m$ can be summarized as follows. Considering Gamma (or Dirichlet) conjugate priors for the parameters

(ξ_{ij}) of the saturated model, when we assume Poisson (or, respectively, multinomial) sampling, we simply generate posterior values from the corresponding posterior distributions. From the posterior values of the (ξ_{ij})'s, using Lemma 1, we generate a posterior sample of the projected values $\boldsymbol{\theta}_m$ and the corresponding distances $\underline{\mathrm{KL}}_m$. The reduced model can be accepted if a descriptive statistic of the posterior distribution of $\underline{\mathrm{KL}}_m$ such as the posterior mean is small enough.

A scaling of the Kullback–Leibler distance may be defined as follows. Let us assume two Poisson models m_1 and m_2 with $y_{ij} \sim \text{Poisson}(\xi^{m_k})$ for $k = 1, 2$. Then, if $\Delta = \xi^{m_2} - \xi^{m_1}$, the Kullback–Leibler distance between these two models is given by

$$\mathrm{KL}(m_1, m_2) = \xi^{m_1} \log \frac{\xi^{m_1}}{\xi^{m_1} + \Delta} + \Delta.$$

If we substitute ξ^{m_1} by the mean cells $\bar{y} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}/(IJ)$, we produce the Kullback–Leibler distance between two Poisson models for departures Δ from the data average of all cell counts. This procedure is an analogue to the one provided by Goutis and Robert(1998) and it is illustrated for our dataset in the next section.

An alternative approach can be based upon the fact that $\underline{\mathrm{KL}}_m$ is monotonically decreasing in m. For successive values of m, we define the indexes

$$\gamma_m = \frac{E(\underline{\mathrm{KL}}_m|\boldsymbol{y})}{E(\underline{\mathrm{KL}}_{m-1}|\boldsymbol{y})}, \quad m = 1, \ldots, M^*.$$

Note that $\gamma_{M^*} = 0$ and $0 \leqslant \gamma_m < 1$. Whenever $\gamma_m = 0$ for $m < M^*$, the order of association is m and $\gamma_q = 0$ for $q > m$. The product

$$g_m = \prod_{q=1}^{m} \gamma_q \tag{3.5}$$

is the proportion of the interaction unexplained by model RC(m) and satisfies the relation $E(\underline{\mathrm{KL}}_m|\boldsymbol{y}) = g_m E(\underline{\mathrm{KL}}_0|\boldsymbol{y})$.

Next in order to evaluate the fit of the assumed model, we propose some indexes taking into account the complexity of the model as well. At first place, it is reasonable to expect that between two models, RC(m_1) and RC(m_2) with d_{m_1} and d_{m_2} number of parameters, respectively, we will prefer RC(m_1) rather than RC(m_2) if the quantity

$$DR(m_1, m_2) = \frac{d_{m_1} E(\underline{\mathrm{KL}}_{m_1}|\boldsymbol{y})}{d_{m_2} E(\underline{\mathrm{KL}}_{m_2}|\boldsymbol{y})} \tag{3.6}$$

is less than one. $DR(m_1, m_2)$ may be viewed as the product of the percentage change in the dimension of the model times the percentage change in the Kullback–Leibler directed divergence between each model and the saturated one. The percentage change of the Kullback–Leibler distance is relevant to the information index of Soofi (1992) that is based on the entropy reduction and it is applied on the conditional logit model. Entropy reduction is also discussed by Gilula and Haberman (1994) in the context of panel data.

The next index we introduce is based on (3.5) and is defined as

$$ID_m^1 = (g_m)^{c_1}\left(\frac{d_{M^*} - d_m}{d_{M^*} - d_0}\right)^{c_2} \tag{3.7}$$

for some positive constants c_1, c_2 that satisfy $c_1 + c_2 = 1$. Note that d_{M^*}, d_m, and d_0 stand for the number of parameters of the saturated, the RC(m) model, and the model of independence, respectively. The role of g_m is explained above and the ratio of the second term in (3.7) is the proportion of the interaction parameters left free by RC(m). Thus, (3.7) is a weighted geometric mean of a measure of model fit (its first term) and a measure of model complexity (its second term). Reasonable choices of c_1 and c_2 are $c_1 = c_2 = 1/2$ (unweighted case, leading to the simple geometric mean) or $c_1 = IJ/(\log n + IJ)$ and $c_2 = 1 - c_1$. In our second proposal, c_1 is a decreasing function of $\log n/(IJ)$. It is easy to verify that $0 < ID_m^1 < 1$ (for $0 < m < M^*$), $ID_0^1 = 1$ and $ID_{M^*}^1 = 0$.

We also use a last index defined as

$$ID_m^2 = 1 - \frac{\sum_{k=1}^{m} E(\phi_k|\boldsymbol{y})^2}{\sum_{k=1}^{M^*} E(\phi_k|\boldsymbol{y})^2}. \tag{3.8}$$

This index is based on the relation $\sum_{k=1}^{M^*}\phi_k^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\lambda_{ij}^{(12)}\right)^2$, which holds when $w_{1i} = w_{2j} = 1$ and it is analogous to a relevant quantity used in correspondence analysis.

Finally, in what follows, we approximate under repeated sampling, the distribution of the posterior mean of $\underline{\mathrm{KL}}_0$ which is the minimum Kullback–Leibler distance for the independence model. For the Poisson sampling, for example, $\underline{\mathrm{KL}}_0$ is given by

$$\underline{\mathrm{KL}}_0 = \sum_{i=1}^{I}\sum_{j=1}^{J}\xi_{ij}\log\xi_{ij} + IJ\bar{\xi}\log\bar{\xi} - J\sum_{i=1}^{I}\bar{\xi}_{i\cdot}\log\bar{\xi}_{i\cdot} - I\sum_{j=1}^{J}\bar{\xi}_{\cdot j}\log\bar{\xi}_{\cdot j},$$

where $\bar{\xi}_{i\cdot} = \sum_{j=1}^{J}\xi_{ij}/J$, $\bar{\xi}_{\cdot j} = \sum_{i=1}^{I}\xi_{ij}/I$ and $\bar{\xi} = \sum_{i=1}^{I}\sum_{j=1}^{J}\xi_{ij}/(IJ)$. The conjugate prior $\xi_{ij} \sim \text{Gamma}(\gamma_{ij}, \delta)$ generates the posterior distributions

$$\begin{aligned}
\xi_{ij}|\boldsymbol{y} &\sim \text{Gamma}(y_{ij} + \gamma_{ij}, \delta + 1)\\
\bar{\xi}|\boldsymbol{y} &\sim \text{Gamma}(\ IJ(\bar{y} + \bar{\gamma}\ ), IJ(\delta + 1)\ ),\\
\bar{\xi}_{i\cdot}|\boldsymbol{y} &\sim \text{Gamma}(\ J(\bar{y}_{i\cdot} + \bar{\gamma}_{i\cdot}\ ), J(\delta + 1)\ ),\\
\bar{\xi}_{\cdot j}|\boldsymbol{y} &\sim \text{Gamma}(\ I(\bar{y}_{\cdot j} + \bar{\gamma}_{\cdot j}\ ), I(\delta + 1)\ ).
\end{aligned}$$

So the posterior expectation of $\underline{\mathrm{KL}}_0$ is

$$\begin{aligned}
E(\underline{\mathrm{KL}}_0|\boldsymbol{y}) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{y_{ij} + \gamma_{ij}}{\delta + 1}\Omega_{ij} + \frac{(I-1)(J-1)}{\delta + 1},\\
\Omega_{ij} &= \Psi(\ y_{ij} + \gamma_{ij}\ ) + \Psi(\ IJ(\bar{y} + \bar{\gamma})\ )\\
&\quad -\Psi(\ J(\bar{y}_{i\cdot} + \bar{\gamma}_{i\cdot})\ ) - \Psi(\ I(\bar{y}_{\cdot j} + \bar{\gamma}_{\cdot j})\ ),
\end{aligned}$$

where $\Psi(x)$ is the digamma function $\Psi(x) = d \log \Gamma(x)/dx$. Its Taylor expansion gives the approximation

$$E(\underline{\mathrm{KL}}_0|\boldsymbol{y}) \approx \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{y_{ij}+\gamma_{ij}}{\delta+1} \log \frac{y_{ij}+\gamma_{ij}}{(\bar{y}_{i\cdot}+\bar{\gamma}_{i\cdot})(\bar{y}_{\cdot j}+\bar{\gamma}_{\cdot j})/(\bar{y}+\bar{\gamma})} + \frac{(I-1)(J-1)}{2\,(\delta+1)}.$$

For small values of γ_{ij} and δ it is simplified to

$$\begin{aligned} E(\underline{\mathrm{KL}}_0|\boldsymbol{y}) &\approx \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log \frac{y_{ij}}{\bar{y}_{i\cdot}\bar{y}_{\cdot j}/\bar{y}} + \frac{(I-1)(J-1)}{2} \\ &\approx -\log \frac{f(\boldsymbol{y}|\hat{\xi}^0_{ij}, 0)}{f(\boldsymbol{y}|\hat{\xi}_{ij}, M^*)} + \frac{(I-1)(J-1)}{2}. \end{aligned}$$

The approximation is valid when δ tends to zero and all the cell counts y_{ij} are much higher than the corresponding prior parameters γ_{ij}.

Therefore, the posterior mean of $\underline{\mathrm{KL}}_0$ is distributed, under repeated sampling, approximately as a $\left(\chi^2_{(I-1)(J-1)}/2 + (I-1)(J-1)/2\right)$ with mean $(I-1)(J-1)$ and variance $(I-1)(J-1)/2$. The interval

$$\left(\frac{1}{2}\chi^2_{(I-1)(J-1),a/2} + \frac{(I-1)(J-1)}{2},\ \frac{1}{2}\chi^2_{(I-1)(J-1),1-a/2} + \frac{(I-1)(J-1)}{2} \right) \tag{3.9}$$

can serve as a basis in our effort to evaluate the magnitude of $E(\underline{\mathrm{KL}}_m|\boldsymbol{y})$.

Remark. The deviance of an RC(m) model is given by

$$D(\boldsymbol{\theta}_m) = 2\left(\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log \frac{y_{ij}}{\xi^m_{ij}} - \mathcal{I}_p(y_{ij} - \xi^m_{ij}) \right).$$

Note that it is equal to twice the KL distance given by (3.2) if the parameters of the saturated model ξ_{ij} are substituted by the data y_{ij}. Popular measures for the evaluation of a model that can be directly computed using the output of the MCMC algorithm are the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, and van der Linde 2002)

$$\mathrm{DIC}_m = 2\overline{D(\boldsymbol{\theta}_m)} - D(\overline{\boldsymbol{\theta}}_m)$$

and the Bayesian versions of AIC and BIC (Brooks 2002)

$$\mathrm{AIC}_m = \overline{D(\boldsymbol{\theta}_m)} + 2d_m \qquad \mathrm{BIC}_m = \overline{D(\boldsymbol{\theta}_m)} + d_m \log(n).$$

The bar denotes the posterior mean value of the corresponding quantity. The drawback of using DIC is that the posterior means cannot be considered as the best estimates in RC(m) models because they do not satisfy the constraints imposed on the parameters.

Table 1. Classification of 5,387 Children in Caithness According to their Hair and Eye Color

	Hair color				
Eye color	*Fair*	*Red*	*Medium*	*Dark*	*Black*
Blue	326	38	241	110	3
Light	688	116	584	188	4
Medium	343	84	909	412	26
Dark	98	48	403	681	85

4. ILLUSTRATIVE EXAMPLE: THE CAITHNESS DATASET

We apply our methods to a well-known example in the association models literature; a 4×5 contingency table of the Caithness children cross-classified by eye and hair color given in Table 1 (Fisher 1940). This dataset was first analyzed using association models by Goodman (1981). The classical likelihood ratio test leads to the conclusion that the order of the underlying association model is equal to $m = 2$.

4.1 BAYESIAN ESTIMATION OF THE RC(*m*) PARAMETERS

Here we present detailed results concerning the posterior distributions of the parameters of RC(1) and RC(2) models estimated using the MCMC algorithm of Section 2. We also present some technical details concerning the RC(3) model. Each row and column score was updated 30 times for each iteration of the rest of the parameters as suggested by Viele and Srinivasan (2000). Initial values for the main effects were taken to be equal to maximum likelihood estimates under the independence model. Initial values for the row and column scores were generated from the uniform distribution on the corresponding unit sphere and $\mathbf{\Phi}_m = \text{diag}(5, 3, 1)$ for $m = 3$. When the dimension the model was less than three the corresponding ϕ_i's were set equal to zero.

For all models we run the MCMC algorithm until the convergence diagnostics of Geweke (1992) and Heidelberger and Welch (1983) were satisfactory. For RC(1) and RC(2) models we have considered 30,000 iterations with lag of 5 and 16 to save storing space and additional 10,000 and 20,000 iterations as a burn-in period, respectively. Plots and summaries of the posterior distributions of the models RC(1) and RC(2) are presented in Table 2 and in Figure 1. The posterior distributions for RC(2) and RC(1) are quite close to the results of Haberman (1995) using the maximum likelihood approach.

For the RC(3) model we considered a sample of 15,000 iterations with lag of 100 iterations and additional 50,000 iterations burn-in. The total number of iterations for the RC(3) model is considerably higher than the other two models due to the complicated posterior distribution of the parameters and the high correlations between them. The posterior distribution of RC(3) was also compared with an MCMC sample (of equal number of iterations) of the saturated log-linear model transformed using singular value decomposition in each iteration. Both samples were tested for convergence using CODA software. Detailed results for the RC(3) and further details for the RC(1) and RC(2) including convergence diagnostics, trace, and ergodic mean plots are available from the authors on request.

4.2 EVALUATION OF THE FIT

We assume that the prior generator for the expected cell frequencies ξ_{ij} is the Gamma$(1, 10^{-4})$. Table 3 displays posterior descriptive measures for the average minimum Kullback–Leibler distance $\underline{\text{AKL}}_m = \underline{\text{KL}}_m/(IJ)$ for all the fitted models as well as the estimates of the ID_m^1 and DR indexes. The ID_m^2 index takes the values .032 and .009 for the RC(1) and RC(2) models, respectively, indicating the superiority of RC(2). According to the approach of Section 3, the Kullback–Leibler directed divergence of the RC(2) from the saturated model is equal to .18 on average which corresponds to the Kullback–Leibler

Table 2. Posterior Summaries for RC(2) and RC(1) Models (MLE: Maximum Likelihood Estimates)

	RC(2)				*RC(1)*			
	MLE	*median*	*mean*	*sd*	*MLE*	*median*	*mean*	*sd*
λ		4.848	4.846	.041	4.831	4.830	.037	
$\lambda_1^{(1)}$	$-.691$	$-.681$	$-.683$	.059	$-.727$	$-.723$	$-.724$	.045
$\lambda_2^{(1)}$	.020	.029	.028	.044	.033	.036	.035	.038
$\lambda_3^{(1)}$	.438	.429	.431	.050	.497	.495	.495	.026
$\lambda_4^{(1)}$	.233	.223	.224	.051	.197	.193	.194	.048
$\lambda_1^{(2)}$	.844	.838	.839	.048	.814	.813	.813	.045
$\lambda_2^{(2)}$	$-.537$	$-.649$	$-.649$	.063	$-.645$	$-.649$	$-.649$	.059
$\lambda_3^{(2)}$	1.324	1.320	1.321	.045	1.383	1.381	1.381	.040
$\lambda_4^{(2)}$	.776	.768	.769	.048	.741	.739	.740	.043
$\lambda_5^{(2)}$	-2.308	-2.271	-2.281	.149	-2.294	-2.280	-2.285	.129
ϕ_1	3.067	3.001	3.020	.236	3.180	3.151	3.156	0.183
ϕ_2	.376	.397	.400	.061				
μ_{11}	.411	.404	.402	.040	$-.421$	$-.420$	$-.420$	.026
μ_{21}	.478	.477	.477	.027	$-.453$	$-.454$	$-.454$	.023
μ_{31}	$-.122$	$-.110$	$-.110$	.048	.095	.095	.095	.023
μ_{41}	$-.767$	$-.771$	$-.770$	.021	.780	.779	.779	.010
μ_{12}	.472	.472	.463	.107				
μ_{22}	$-.030$	$-.017$	$-.012$	.118				
μ_{32}	$-.804$	$-.806$	$-.797$	.054				
μ_{42}	.362	.349	.345	.059				
ν_{11}	.574	.579	.579	.029	$-.594$	$-.595$	$-.595$	.026
ν_{21}	.279	.277	.277	.034	$-.270$	$-.269$	$-.268$	.034
ν_{31}	.121	.121	.120	.021	$-.117$	$-.117$	$-.116$	.019
ν_{41}	$-.260$	$-.274$	$-.273$	.045	.290	.294	.294	.034
ν_{51}	$-.714$	$-.704$	$-.703$	.033	.691	.687	.685	.027
ν_{12}	.271	.293	.280	.174				
ν_{22}	.148	.129	.124	.213				
ν_{32}	$-.838$	$-.809$	$-.781$	.098				
ν_{42}	.448	.369	.355	.237				
ν_{52}	$-.030$	.014	.021	.170				

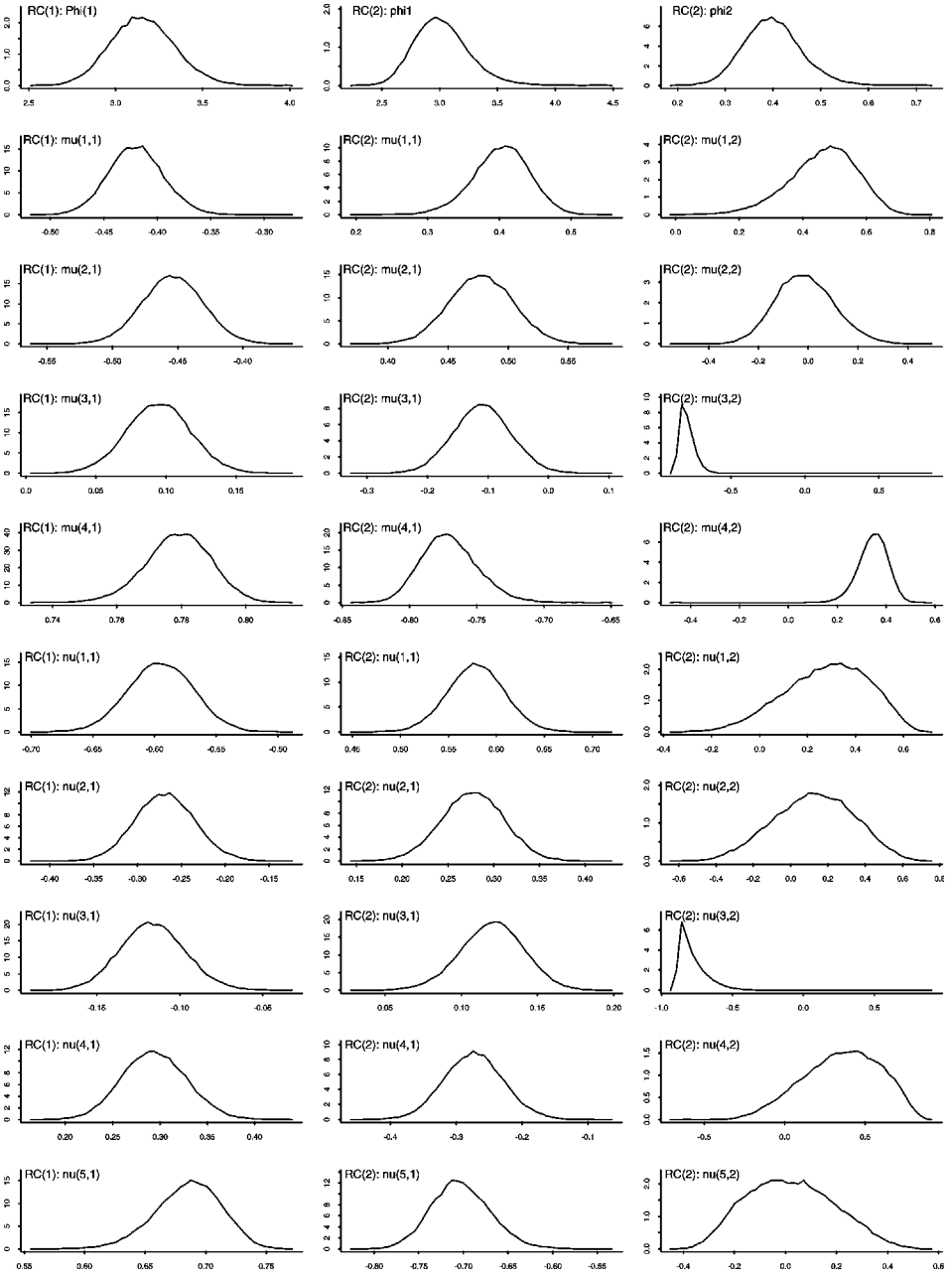

Figure 1. Posterior Densities of the Interaction Parameters of RC(1) and RC(2) Models. Results have been generated from samples of 30,000 iterations with lag 5 of 16 iterations for RC(1) and RC(2), respectively. Additional 10,000 and 20,000 iterations were discarded as burn-in period.

Table 3. Posterior Values of $\underline{\text{AKL}}_m = \underline{\text{KL}}_m/(\text{IJ})$ and Estimates of Proposed Indexes

	Posterior AKL				$\widehat{ID}_m^1$		
Model	*Mean*	*Median*	*SD*	$\hat{\gamma}_m$	$c_1 = .5$	$c_1 = \frac{IJ}{\log n + IJ}$	$\widehat{DR}(m_1, m_2)$
Constant	124.49	124.48	3.11				
Indep.($m = 0$)	30.67	30.65	1.69		1.000	1.000	$\widehat{\text{DR}}(0,1) = 10.02$
RC(1) ($m = 1$)	1.75	1.73	.42	.057	.169	.110	$\widehat{\text{DR}}(1,2) = 7.77$
RC(2) ($m = 2$)	.18	.15	.12	.100	.069	.026	$\widehat{\text{DR}}(0,2) = 77.89$

directed divergence between two Poisson distributions with means $\xi^{m_1} = \bar{y} = 269.35$ and $\xi^{m_2} = \xi^{m_1} + \Delta$ with $\Delta = \pm 10$ (approximately equal to $\pm 3.7\%$ of the expected value of cell counts). The corresponding yardstick values for the RC(1), independence and constant models are much higher and are given in Table 4. For calibration of $\underline{\text{AKL}}_m$ see also Figure 2. The RC(2) model is also supported against RC(1) model by $\hat{DR}(1,2) = 7.77 > 1$. Thus, all diagnostics indicate that the adequate association model for this dataset is the RC(2).

Moreover, we use the calibration approach of McCulloch and Rossi (1993) to produce Table 5. Using this table the divergence of the RC(2) model is equivalent to predicting a Bernoulli distribution with actual $p = .5$ using $p = .77$, a Poisson distribution with actual $\lambda = 1.0$ using $\lambda = .52$ or 1.71 and equivalent to a Normal distribution $N(\mu, \sigma^2)$ using a mean which is larger or smaller by .84 standard deviations (see also Table 6).

Finally, we compare the posterior means of the models under consideration with the bounds of the interval (3.9); see Table 6. The posterior mean of the $\underline{\text{KL}}_{\text{RC}(1)}$ is equal to 35 ($= 1.75 \times 20$) which is much higher than the 99% upper bound 20.15. Therefore, we have a clear evidence that RC(1) model should be rejected because its Kullback–Leibler projection is much worse than the corresponding expected projection when the model of independence is true. On the other hand, for the RC(2) model $E(\underline{\text{KL}}_{\text{RC}(2)}|\boldsymbol{y}) = .18 \times 20 = 3.6$ is less than the 99% lower bound 7.54 hence no evidence is induced against model RC(2) since its Kullback–Leibler projection is much lower than the corresponding expected projection when the model of independence is true.

We conclude our illustration calculating the Bayesian measures of deviance, AIC, BIC, and DIC. The latter was computed using both parameterizations $\boldsymbol{\theta}_m$ and $\boldsymbol{\xi}^m$. Measures for the full model have been also calculated using the MCMC output of the saturated log-linear model given by (1.1) transformed to RC(3) model using singular value decomposition of the interaction terms in each iteration. Results were found identical with the results of the

Table 4. Corresponding Yardstick Comparisons of Poisson(ξ^{m_1}) Versus Poisson(ξ^{m_2}) with $\xi^{m_1} = \bar{y} = 269.35$ and $\xi^{m_2} = \xi^{m_1} + \Delta$

Model	ξ^{m_2}		Δ		$100 \times (\Delta/\xi^{m_1})$	
Constant	77.3	650.3	−192	381	−71.3	141.5
Indep. ($m = 0$)	161.3	419.3	−108	150	−40.1	55.7
RC(1)	240.3	301.3	−29	32	−10.8	11.9
RC(2)	259.3	279.3	−10	10	−3.7	3.7

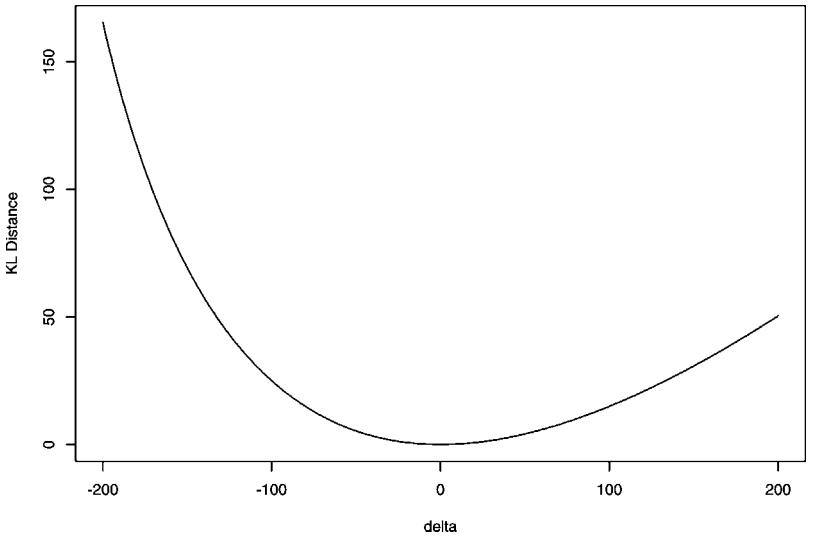

Figure 2. Kullback–Leibler Distance for Departures Δ from the mean of data cells $\bar{y} = 269.35$

RC(3) model estimated using our algorithm proposed in Section 2.

The results are provided in Table 7 and Figures 3 and 4. BIC clearly supports RC(2) model as the KL approach described above. AIC statistics support weakly the RC(2) model (except when we calculate it using the posterior mean deviance from the MCMC output of method proposed in section 2). Generally, the posterior distributions of AIC for RC(2) and RC(3) models are quite close; see Figure 3. For this reason we cannot clearly discriminate between RC(2) and RC(3) models using AIC. Finally, although the two versions of DIC give different results, they both support the RC(3) model. If we use the second version of DIC (last column of Table 7), which seems to give more reliable results, we see that RC(2) and RC(3) are close.

5. DISCUSSION

In this article we have implemented Bayesian inference for RC(m) models focusing on the estimation of the posterior distributions of the model parameters and on the evaluation

Table 5. Calibrated Values that Correspond to the Posterior Mean of $\underline{\text{AKL}}_m$. In the comparison of the Normal distributions (column 4) σ is assumed to be common.

	Bernoulli p *vs. .5*	*Poisson* λ *vs. 1*	*Normal* $\Delta = \frac{\mu_1 - \mu_2}{\sigma}$
Constant	1.000	149.90	23.99
Independence	1.000	35.23	11.07
RC(1)	.992	4.18	2.60
RC(2)	.772	.52, 1.71	.84

Table 6. Intervals for Evaluating the Plausibility of KL distances. UB: Upper Bound, LB: Lower Bound.

	Independence Model			
α	$\chi^2_{a/2}$	$\chi^2_{1-a/2}$	*LB*	*UB*
.10	5.23	21.03	8.61	16.51
.05	4.40	23.34	8.20	17.67
.01	3.07	28.30	7.54	20.15

of the fit of such models. The specification and elicitation of prior information for the parameters of the RC(m) model is not an easy task. A prior distribution may be constructed using imaginary data (Ibrahim, Chen, and Shao 2000; Ibrahim and Chen 2002). Under this approach, we define the prior distribution to have a form similar to a fraction of the likelihood function

$$f(\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{\Phi}_m, \boldsymbol{M}_m, \boldsymbol{N}_m | \boldsymbol{y}^*) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} f(y^*_{ij} | \lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{\Phi}_m, \boldsymbol{M}_m, \boldsymbol{N}_m)^{c_{ij}} \tag{5.1}$$

assuming imaginary data $\boldsymbol{y}^*$; the prior parameters $c_{ij} \geq 0$ play the role of dispersion for each imaginary data point y^*_{ij}. We may interpret c_{ij} as a measure of "how much we believe" the corresponding imaginary data and hence how much weight we wish to attribute in each prior data. If we wish to give to the prior weight equal to one data point, we set $c_{ij} = c = 1/n$. To gain a feeling about the location and the shape of the induced marginal prior distributions we could either employ MCMC using the imaginary data or, when $c_{ij} = c$, we could estimate the prior modes using the algorithms of Becker (1990) and Haberman (1995) substituting the actual with the imaginary data. Note that the parameter c influences only the prior variances and not the posterior means.

Using a similar argument to set a prior distribution for the expected cell frequencies of the saturated model ξ_{ij}, leads, in the case of Poisson sampling, to the conjugate gamma prior distribution Gamma$(\gamma_{ij}, \delta_{ij})$ with $\gamma_{ij} = c_{ij} y^*_{ij} + 1$ and $\delta_{ij} = c_{ij}$. The prior Gamma$(1, 10^{-4})$ that was used in the example (Section 4) can be interpreted as a set of imaginary data with $y_{ij} = 1$ weighted by $c_{ij} = 10^{-4}$. In the case of the multinomial model, any set of imaginary data $\boldsymbol{y}^*$ weighted by c_{ij} results to a Dirichlet conjugate prior

Table 7. Posterior Values of the Deviance, AIC, BIC, and DIC

	Deviance				*AIC*		*BIC*		*DIC*	
Model	*Min*	*Mean*	$D(\bar{\boldsymbol{\theta}}_m)$	$D(\bar{\boldsymbol{\xi}}^m)$	*Min*	*Mean*	*Min*	*Mean*	$D(\bar{\boldsymbol{\theta}}_m)$	$D(\bar{\boldsymbol{\xi}}^m)$
Constant	4966.0	4967.2	4966.2	4966.2	4968.0	4969.2	4974.6	4975.8	4968.2	4968.2
Indep.($m = 0$)	1219.0	1226.4	1218.3	1218.4	1235.0	1242.4	1287.7	1295.1	1234.5	1234.4
RC(1) ($m = 1$)	65.8	78.6	64.7	64.4	93.8	106.6	186.1	198.9	92.5	92.8
RC(2) ($m = 2$)	7.2	23.2	5.7	5.4	43.2	59.2	161.9	177.9	40.7	41.0
RC(3) ($m = 3$)	4.5	18.9	36.5	.5	44.5	58.9	176.4	190.8	1.3	37.3
Full ($m = 3$)[1]	3.4	20.1	36.5	.0	43.4	60.1	175.3	191.9	3.7	40.2

[1] Results are based on MCMC output of the saturated log-linear model and then transforming the interaction terms using singular value decomposition.

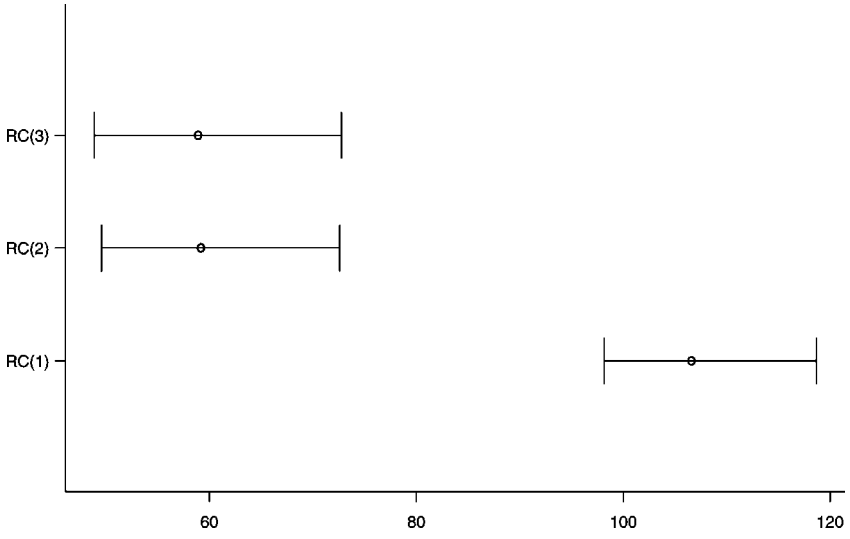

Figure 3. 95% credible intervals for AIC.

distribution $\mathcal{DI}_{IJ-1}(\boldsymbol{\zeta})$ with parameters $\zeta_{ij} = c_{ij}y^*_{ij} + 1$.

A referee asked how available prior information on the interactions could be quantified as a prior for the parameters of the RC(m) model. Starting from a prior distribution for the expected cell frequencies of the saturated model ξ_{ij}, we can generate Monte Carlo samples for them and compute the mean $\bar{\xi_{ij}}$ and the variance $s^2_{\xi_{ij}}$ of each ξ_{ij}. Matching them with the mean and the variance of the fractional prior distribution Gamma$(c_{ij}y^*_{ij} + 1, c_{ij})$ of the Poisson sampling, corresponds to using as imaginary data

$$y^*_{ij} = \min\left\{0, \frac{s^2_{\xi_{ij}}}{\bar{\xi_{ij}}}\left(\frac{\bar{\xi_{ij}}^2}{s^2_{\xi_{ij}}} - 1\right)\right\}$$

with prior weights $c_{ij} = \bar{\xi_{ij}}/s^2_{\xi_{ij}}$.

It is worth mentioning that the estimation of the posterior mode using the fractional prior (5.1), is similar in terms of computational effort to the estimation of maximum likelihood estimates. Under the assumption of Poisson sampling, using $c_{ij} = c$, the corresponding posterior distribution is given by

$$\begin{aligned} f(\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{\Phi}_m, \boldsymbol{M}_m, \boldsymbol{N}_m | \boldsymbol{y}, \boldsymbol{y}^*) \\ \propto f(\boldsymbol{y}|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{\Phi}_m, \boldsymbol{M}_m, \boldsymbol{N}_m) f(\boldsymbol{y}^*|\lambda, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \boldsymbol{\Phi}_m, \boldsymbol{M}_m, \boldsymbol{N}_m)^c \\ \propto \exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J}\xi_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{y_{ij} + cy^*_{ij}}{1+c}\log\xi_{ij}\right)^{1+c}. \end{aligned}$$

The posterior mode can then be estimated by substituting the actual data by the weighted mean $\bar{y}_{ij} = (y_{ij} + cy^*_{ij})/(1 + c)$ in the algorithm of Becker (1990) or Haberman (1995). The multinomial case is treated similarly.

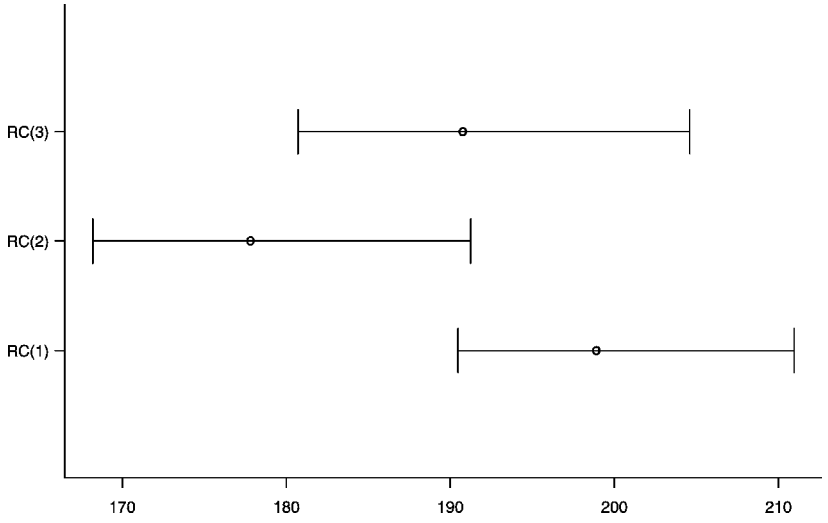

Figure 4. 95% credible intervals for BIC

Touching on the extension of the MCMC methodology over parameter spaces of different dimension using either reversible jump MCMC (Green 1995) or other related methods (see for example in Dellaportas, Forster, and Ntzoufras 2002), we express some of our concerns with respect to different methods. Although Gibbs sampling methods used for variable selection seem to be directly applicable in RC(m) models, there are still some important tasks that should be considered. In the stochastic search variable selection (George and McCulloch 1993) the specification of sensible "small" prior distributions is not a straightforward task especially in this case that the prior distributions on ϕ_k cannot be a mixture of Normal distributions. Furthermore, the Gibbs sampler of Kuo and Mallick (1998) might not be very efficient (see for discussion and drawbacks Dellaportas, Forster, and Ntzoufras 2000), while the Gibbs variable selection of Dellaportas et al. (2002) involves careful specification of pseudo-priors that might be difficult within the formulation of RC(m) models due to the complicated nature of the parameters. For all the above reasons, the construction of an efficient MCMC algorithm for estimating the posterior distribution of the dimension m of an RC(m) model is a challenging task and requires further research.

Finally, we conclude our discussion by underlying that the methodology introduced in this article can serve as the concrete base for the constructing of full Bayesian analysis in RC(m) models. Further work on this aspect may include extensions of the ideas concerning the incorporation of prior information, evaluation of posterior model probabilities and implementation of Bayesian model averaging. Although, the MCMC methodology proposed in Section 2 is much more computationally demanding than the frequentist approach, the two approaches cannot be directly compared because the first approach explores the whole posterior distribution providing more insight concerning the distribution of the parameters, although the latter only calculates the maximum of the likelihood. What we may compare,

in terms of computational effort, is the frequentist approach with the method proposed to calculate the posterior mode as described in this section. The two procedures will be computationally identical because the same algorithms will be used with slightly different values as input data.

APPENDIX

Proof of Lemma 1: The minimization problem is equivalent for both sampling schemes, because they belong to the exponential family. We provide details of the proof considering the multinomial sampling. We write the density function under the multinomial sampling in the form

$$f(\boldsymbol{y}|\boldsymbol{\xi}) = h(\boldsymbol{y}) \exp\left(\sum_{(i,j)\neq(I,J)} Q(\xi_{ij}) y_{ij} - c(\boldsymbol{\xi}) \right), \tag{A.1}$$

where $Q(\xi_{ij}) = \log(\xi_{ij}/\xi_{IJ}) = \log\{\xi_{ij}/(n - \sum_{(i,j)\neq(I,J)} \xi_{ij})\}$, for $i = 1, \ldots, I$, $j = 1, \ldots, J$, $(i,j) \neq (I,J)$, $h(\boldsymbol{y}) = n! n^{-n}/(y_{11}! y_{12}! \ldots y_{IJ}!)$, $\boldsymbol{\xi} = (\xi_{ij})$ and $c(\boldsymbol{\xi}) = -n \log(n - \sum_{(i,j)\neq(I,J)} \xi_{ij})$. A reformulation of the density in terms of the natural parameter vector $\boldsymbol{\psi}$ with $\psi_{ij} = Q(\xi_{ij})$, is

$$f(\boldsymbol{y}|\boldsymbol{\psi}) = h(\boldsymbol{y}) \exp\left(\sum_{(i,j)\neq(I,J)} \psi_{ij} y_{ij} - b(\boldsymbol{\psi}) \right), \tag{A.2}$$

with $b(\boldsymbol{\psi}) = n \log\{1 + \sum_{(i,j)\neq(I,J)} \exp(\psi_{ij})\} - n \log(n)$. The Kullback–Leibler distance between the sampling distributions under the true underlying model and the assumed RC(m), in terms of ψ_{ij}, is given by

$$\mathrm{KL}[f(\boldsymbol{y}|\boldsymbol{\psi}), f(\boldsymbol{y}|\boldsymbol{\psi}_m)] = \sum_{i=1}^{I} \sum_{j=1}^{J} f(y_{ij}|\boldsymbol{\psi}) \log\left(\frac{\exp\{\psi_{ij} y_{ij} - b(\psi_{ij})\}}{\exp\{\psi_{ij}^m y_{ij} - b(\psi_{ij}^m)\}} \right),$$

which can equivalently be expressed as

$$\mathrm{KL}[f(\boldsymbol{y}|\boldsymbol{\psi}), f(\boldsymbol{y}|\boldsymbol{\psi}_m)] = E_{\boldsymbol{y}|\boldsymbol{\psi}}[(\psi_{ij} - \psi_{ij}^m) y_{ij} - b(\psi_{ij}) + b(\psi_{ij}^m)].$$

Using the fact that $E_{\boldsymbol{y}|\boldsymbol{\psi}}(y_{ij}) = \xi_{ij} = b'(\psi_{ij})$, we conclude that

$$\mathrm{KL}[f(\boldsymbol{y}|\boldsymbol{\psi}), f(\boldsymbol{y}|\boldsymbol{\psi}_m)] = \sum_{i=1}^{I} \sum_{j=1}^{J} \left\{ (\psi_{ij} - \psi_{ij}^m) \xi_{ij} - b(\psi_{ij}) + b(\psi_{ij}^m) \right\}. \tag{A.3}$$

In order to minimize the desired Kullback–Leibler distance we equate to zero the first derivatives of (1.3) with respect to the components of $\boldsymbol{\theta}_m$, $(\lambda_i^{(1)}, \lambda_j^{(2)}, \mu_{ik}^m, \nu_{jk}^m, \phi_k^m)$ $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, M$. Note that the parameter λ of (1.3) is produced through the linear transformation of the $\lambda_i^{(1)}$'s and $\lambda_j^{(2)}$'s in order to center them around zero. □

Proof of Lemma 2. Denote by ξ_{ij}^0 the expected cell frequencies under the model of independence. Substituting ξ_{ij}^0, ξ_{ij}, and ξ_{ij}^m in the log-expressions of $\mathrm{KL}(0, M^*)$ and $\mathrm{KL}(0, m)$, respectively, by (1.3) and using marginal weights for the row and column scores we conclude, after some algebra, that $\mathrm{KL}(0, M^*) = n(\lambda_0 - \lambda) + \sum_i \xi_{i\cdot}(\lambda_{i(0)}^{(1)} - (\lambda_{i(M^*)}^{(1)}) + \sum_j \xi_{\cdot j}(\lambda_{j(0)}^{(2)} - (\lambda_{j(M^*)}^{(2)})$. An analogous expression is derived for the $\mathrm{KL}(0, m)$. It can also be proved that $\mathrm{KL}(m, 0) + \mathrm{KL}(0, m) = \sum_{k=1}^m \phi_k^m \rho_k^m$ for $m = 1, \ldots, M^*$, where $\rho_k^m = \sum_{i,j} \xi_{ij}^m \mu_{ik}^m \nu_{jk}^m$ is the weighted by $\boldsymbol{\pi} = (\xi_{ij}/n)$ correlation between the row and column scores of the kth dimension of the association (Eshima, Tabata, and Tsujitani 2001). Combining the above relations and noting that the marginals $\xi_{i\cdot}$ and $\xi_{\cdot j}$ remain the same under any RC(m) model ($m \geq 0$), we conclude that

$$\mathrm{KL}(M^*, m) = \mathrm{KL}(M^*, 0) - \mathrm{KL}(m, 0) + \sum_{k=1}^{m} \phi_k^m \left(\sum_{i=1}^{I} \sum_{j=1}^{J} (\xi_{ij}^m - \xi_{ij}) \mu_{ik}^m \nu_{jk}^m \right), \quad \text{(A.4)}$$

which holds in general for $m \geq 1$. When the expected frequencies ξ_{ij}^m are estimated through the Kullback–Leibler projection, then by Lemma 1, the sum in (A.4) is equal to zero. □

[Received March 2003. Revised January 2004.]

REFERENCES

Agresti, A., and Chuang, C. (1989), "Model-Based Bayesian Methods for Estimating Cell Proportions in Cross-Classification Tables Having Ordered Categories," *Computational Statistics and Data Analysis*, 7, 245–258.

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.

Becker, M. (1990), "Maximum Likelihood Estimation of the RC(M) Association Model," *Applied Statistics*, 39, 152–167.

Brooks, S. P. (2002), Discussion of "Bayesian Measures of Model Complexity and Fit" by Spiegelhalter, Best, Carlin, and van der Linde, *Journal of the Royal Statistical Society*, Ser. B, 64, 616–618.

Chipman, H., George, E. I., and McCulloch, R. E. (2001), "The Practical Implementation of Bayesian Model Selection," *IMS Lecture Notes—Monograph Series 38: Model Selection*, 67–116.

Chuang, C. (1982), "Empirical Bayes Methods for a Two-Way Multiplicative-Interaction Model," *Communications in Statistics: Theory and Methods*, 11, 2977–2989.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2000), "Bayesian Variable Selection Using the Gibbs Sampler," *Generalized Linear Models: A Bayesian Perspective* eds. D.K. Dey, S. Ghosh, and B. Mallick, New York: Marcel Dekker, pp. 271–286.

——— (2002), "On Bayesian Model and Variable Selection Using MCMC," *Statistics and Computing*, 12, 27–36.

Dupuis, J. A. (1997), "Bayesian Test of Homogeneity for Markov Chains," *Statistics and Probability Letters*, 31, 333–338.

Eshima, N., Tabata, M., and Tsujitani, M. (2001), "Property of the RC(M) Association Model and a Summary Measure of Association in the Contingency Table," *Journal of the Japananese Statistical Society*, 31, 15–26.

Evans, M., Gilula, Z., and Guttman, I. (1993), "Computational Issues in the Bayesian Analysis of Categorical Data: Log-linear and Goodman's RC model," *Statistica Sinica*, 3, 391–406.

Fisher, R. A. (1940), "The Precision of Discriminant Functions," *Annals of Eugenics*, 10, 422–429.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: University Press, pp. 169–193.

Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.

Gilula, Z., and Haberman, S. J. (1994), "Conditional Log-Linear Models for Analyzing Categorical Panel Data," *Journal of the American Statistical Association*, 89, 645–656.

Gokhale, D., and Kullback, S. (1978), *The Information in Contingency Tables*, New York: Marcel Dekker.

Goodman, L. A. (1981), "Association Models and Canonical Correlation in the Analysis of Cross-Classifications Having Ordered Categories," *Journal of the American Statistical Association*, 76, 320–334.

——— (1985), "The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models and Asymmetry Models for Contingency Tables With or Without Missing Entries," *The Annals of Statistics*, 13, 10–69.

——— (1991), "Measures, Models and Graphical Displays in the Analysis of Cross-classified Data" (with discussion), *Journal of the American Statistical Association*, 86, 1085–1138.

Goutis, C., and Robert, C. (1998), "Model Choice in Generalized Linear Models: A Bayesian Approach via Kullback-Leibler Projections," *Biometrika*, 85, 29–37.

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

Haberman, S. J. (1995), "Computation of Maximum Likelihood Estimates in Association Models," *Journal of the American Statistical Association*, 90, 1438–1446.

Heidelberger, P., and Welch, P. (1983), "Simulation Run Length Control in the Presence of an Initial Transient," *Operations Research*, 31, 1109–1144.

Ibrahim, J., and Chen, M.-H. (2002), "The Relationship Between the Power Prior and Hierarchical Models," *7th Valencia Meeting on Bayesian Statistics*.

Ibrahim, J., Chen, M.-H., and Shao, Q.-M. (2000), "Power Prior Distributions for Generalized Linear Models," *Journal of Statistical Planning and Inference*, 84, 121–137.

Kuo, L., and Mallick, B. (1998), "Variable Selection for Regression Models," *Sankya*, B, 60, 65–81.

Lopes, H. F. (2002), "Bayesian Model Selection," Technical Report, Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil. Available at http://acd.ufrj.br/~hedibert.

McCulloch, R. E., and Rossi, P. E. (1993), "Bayes Factors for Nonlinear Hypotheses and Likelihood Distributions," *Biometrika*, 79, 663–676.

Soofi, E. S. (1992), "A Generalizable Formulation of Conditional Logit With Diagnostics," *Journal of the American Statistical Association*, 87, 812–816.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996), *BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual*, MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Viele, K., and Srinivasan, C. (2000), "Parsimonious Estimation of Multiplicative Interaction in Analysis of Variance Using Kullback-Leibler Information," *Journal of Statistical Planning and Inference*, 84, 201–219.

Vines, S. K., Gilks, W. R., and Wild, P. (1996), "Fitting Bayesian Multiple Random Effects Models," *Statistics and Computing*, 6, 337–346.