Statistica Sinica 3(1993), 391-406

COMPUTATIONAL ISSUES IN THE BAYESIAN ANALYSIS OF CATEGORICAL DATA: LOG-LINEAR AND GOODMAN'S RC MODEL

Mike Evans, Zvi Gilula* and Irwin Guttman

*University of Toronto and Hebrew University**

Abstract: The Baysian analysis of loglinear models requires the evaluation of high-dimensional integrals. Such an evaluation is frequently computationally prohibitive even with modern computers. We provide a parameterization of the loglinear model which renders these integrations amenable to the numerical methods of adaptive important sampling. This approach is applied in the analysis of two-way contingency tables using Goodman's RC model. We base the analysis on the full posterior distribution for the loglinear model and obtain the posterior distribution of a goodness-of-fit measure for Goodman's RC model.

Key words and phrases: Adaptive importance sampling, categorical data, Goodman's RC model, loglinear models, singular value decomposition.

1. Introduction

Bayesian analysis (of categorical data) may involve the evaluation of high-dimensional integrals which can be (see Evans, Gilula and Guttman (1989)) computationally prohibitive, even for modern computers. Growing interest in the analysis of categorical data by Bayesian methods (as is evident from Leonard (1975), Leonard, Hsu and Tsui (1989), Agresti and Chuang (1989) and Epstein and Fienberg (1990), to mention a few) provides strong motivation for the development of relatively simple evaluation methods to overcome the above mentioned computational difficulties. The derivation of such methods is the purpose of this paper. Effective computational approaches are provided here for the Bayesian analysis of log-linear models with prior structure as in random effects models. A method is proposed for analyzing submodels which prescribe a specific structure for the interaction terms, based on the full posterior distribution of the log-linear model. This method has some advantages over the more common method of Bayesian analysis where a prior distribution is placed on the parameters of the submodels. Our method is then implemented on the important Goodman's RC model (Goodman (1979, 1981, 1985)).

The approach to the computational issues here is similar in spirit to that used in Evans, Gilula and Guttman (1989), which is based on a useful reparametrization of the underlying model and the use of importance sampling. Again we reparametrize the basic model so that importance sampling is easily implemented. The importance sampling approach is seen to be necessary, not only for situations where a normal approximation to the posterior may perform poorly, but generally for the computation of the posterior distributions of the parameters associated with Goodman's RC model and the assessment of fit.

2. The Log-Linear Model

Consider the case of a 2-way table given by categorical variables A and B and note that everything we say can be generalized to the k-way table in a straightforward way. Suppose then that A takes I values, B takes J values, p_{ij} is the probability of a response being categorized in the (i,j)-th cell, $\{p_{i\cdot}\}$, $1 \le i \le I$ and $\{p_{\cdot j}\}$, $1 \le j \le J$ denote the marginal distributions of A and B, respectively. Hence we are assuming multinomial sampling, but note that with minor adjustments all our discussion applies to the product multinomial case as well. The notations used hereafter are similar to those used in Haberman (1974 a,b).

The problem of interest is to offer parametric structures for the association between the variables A and B. Such parametric structures should be both relatively easy to interpret and parsimonious at the same time. For instance, association between variables in Table 1, which is thoroughly analyzed in Section 6, can be described by a *saturated* model having 23 parameters. The model we offer for this table (the RC model) contains much fewer parameters (11), which are shown in Section 3 to have an attractive interpretation.

We first describe, in some necessary length, the way of writing the general saturated model for a two-way contingency table. The meaning of the parameters of such a model are addressed, and then we proceed to consider the parsimonious and interpretable RC model.

Now let $\boldsymbol{p} = (p_{11}, p_{12}, \ldots, p_{IJ})'$, $C_A \in \mathbf{R}^{I\times I}$, $C_B \in \mathbf{R}^{J\times J}$ be arbitrary orthogonal matrices whose first columns are constant, i.e., contrast matrices, and put $C = C_A \otimes C_B$ where $\otimes$ denotes the Kronecker product. We shall denote the columns of C using double subscripts, applying the same ordering as that used in $\boldsymbol{p}$ above. We then have $\boldsymbol{c}_{ij} = \boldsymbol{c}_{iA} \otimes \boldsymbol{c}_{jB}$ where $\boldsymbol{c}_{iA}$, $\boldsymbol{c}_{jB}$ are the ith and jth columns of C_A and C_B respectively. Putting $D_A = (\boldsymbol{c}_{2A} \ldots \boldsymbol{c}_{IA})$ and $D_B = (\boldsymbol{c}_{2B} \ldots \boldsymbol{c}_{JB})$ we write the loglinear model as

$$\ln \boldsymbol{p} = C\boldsymbol{\alpha} = \alpha_{11}\boldsymbol{c}_{1A} \otimes \boldsymbol{c}_{1B} + (D_A\boldsymbol{\alpha}_A) \otimes \boldsymbol{c}_{1B} + \boldsymbol{c}_{1A} \otimes (D_B\boldsymbol{\alpha}_B) + D_A \otimes D_B\boldsymbol{\alpha}_{AB} \quad (1)$$

where $\boldsymbol{\alpha}_A = (\alpha_{21}, \ldots, \alpha_{I1})'$, $\boldsymbol{\alpha}_B = (\alpha_{12}, \ldots, \alpha_{1J})'$ and $\boldsymbol{\alpha}_{AB} = (\alpha_{22}, \alpha_{23}, \ldots)'$.

The parameterization in terms of the α_{ij} is necessary for the computational techniques used in Section 5. In particular α_{11} isolates the single degree of freedom lost in the log-linear model due to the constraint on the p_{ij}. The remaining α_{ij} are free to vary (mathematically) independently on $\mathbf{R}$ (see the discussion after Equation (6)). This is a useful property for importance sampling, as there is no need to build any mathematical dependencies into the algorithm, and it is not shared, to the same extent, by some of the more standard methods of parameterizing this model. This remark applies with even greater force for multidimensional tables.

An alternative parameterization is also useful for interpretative purposes; namely

$$\ln \boldsymbol{p} = \lambda \mathbf{c}_{1A} \otimes \mathbf{c}_{1B} + \phi_A \boldsymbol{\lambda}_A \otimes \mathbf{c}_{1B} + \phi_B \mathbf{c}_{1A} \otimes \boldsymbol{\lambda}_B + \phi_{AB} \boldsymbol{\lambda}_{AB} \tag{2}$$

where $\lambda = \alpha_{11}$ and

$$\phi_A = \left\| \sum_{i=2}^{I} \alpha_{i1} \mathbf{c}_{iA} \right\| = \left[\sum_{i=2}^{I} \alpha_{i1}^2 \right]^{1/2}, \quad \boldsymbol{\lambda}_A = \phi_A^{-1} \sum_{i=2}^{I} \alpha_{i1} \mathbf{c}_{iA}, \tag{3}$$

$$\phi_B = \left\| \sum_{i=2}^{I} \alpha_{ij} \mathbf{c}_{iB} \right\| = \left[\sum_{i=2}^{I} \alpha_{ij}^2 \right]^{1/2}, \quad \boldsymbol{\lambda}_B = \phi_B^{-1} \sum_{i=2}^{I} \alpha_{ij} \mathbf{c}_{jB}, \tag{4}$$

and

$$\phi_{AB} = \left\| \sum_{i=2}^{I} \alpha_{ij} \mathbf{c}_{ij} \right\| = \left[\sum_{i=2}^{I} \sum_{j=2}^{J} \alpha_{ij}^2 \right]^{1/2}, \quad \boldsymbol{\lambda}_{AB} = \phi_{AB}^{-1} \sum_{i=2}^{I} \sum_{j=2}^{J} \alpha_{ij} \mathbf{c}_{ij}. \tag{5}$$

These quantities have useful interpretations.

Since $\mathbf{1}'\boldsymbol{p} = 1$ we have for given ϕ_A, $\boldsymbol{\lambda}_A$, ϕ_B, $\boldsymbol{\lambda}_B$, ϕ_{AB}, $\boldsymbol{\lambda}_{AB}$

$$\lambda = -\sqrt{IJ} \ln \left[\frac{1}{IJ} \sum_i \sum_j \exp \left\{ \phi_A \boldsymbol{\lambda}_{Ai} / \sqrt{J} + \phi_B \boldsymbol{\lambda}_{Bj} / \sqrt{I} + \phi_{AB} \boldsymbol{\lambda}_{ABij} \right\} \right] \tag{6}$$

and the vectors $\phi_A \boldsymbol{\lambda}_A \in \mathcal{L}\{\mathbf{c}_{A2}, \ldots, \mathbf{c}_{AI}\}$, $\phi_B \boldsymbol{\lambda}_B \in \mathcal{L}\{\mathbf{c}_{B2}, \ldots, \mathbf{c}_{BJ}\}$ and $\phi_{AB} \boldsymbol{\lambda}_{AB} \in \mathcal{L}\{\mathbf{c}_{ij} \mid i, j \neq 1\}$ are free in their respective spaces. Hence the α_{ij} are free in $\mathbf{R}$ for $(i, j) \neq (1, 1)$.

The ϕ and λ parameters represent quantities of inferential interest in the analysis. For example, A and B are statistically independent if and only if $\ln p_{ij} = \delta_i + \epsilon_j$ for some δ_i, ϵ_j, and for all i and j. This is easily seen to be true if and only if $\alpha_{ij} = 0$ whenever $i \neq 1$ and $j \neq 1$. Hence A and B are statistically independent if and only if $\phi_{AB} = 0$. Clearly ϕ_{AB} is a measure of association while the λ_{ABij} give information as to the form of the nonindependence when

$\phi_{AB} \neq 0$. If $\phi_{AB} = 0$, then $\phi_A = 0$ if and only if $p_{1\cdot} = \cdots = p_{I\cdot}$ and hence ϕ_A is a measure of homogeneity with the λ_{Ai}, giving information as to the form of the nonhomogeneity when $\phi_A \neq 0$. Similar arguments hold for ϕ_B and λ_{Bj}.

Note that the ϕ and λ parameters are independent of the choice of C_A and C_B. For, if C_A^*, C_B^* are two different contrast matrices, then $\mathcal{L}(C_A \otimes \mathbf{1}_J) = \mathcal{L}(C_A^* \otimes \mathbf{1}_J)$, $\mathcal{L}(\mathbf{1}_I \otimes C_B) = \mathcal{L}(\mathbf{1}_I \otimes C_B^*)$; hence $\mathcal{L}(D_A \otimes D_B) = \mathcal{L}^{\perp}(C_A \otimes \mathbf{1}_J, \mathbf{1}_I \otimes C_B) = \mathcal{L}^{\perp}(C_A^* \otimes \mathbf{1}_J, \mathbf{1}_I \otimes C_B^*) = \mathcal{L}(D_A^* \otimes D_B^*)$. Then, for example, $\phi_A \boldsymbol{\lambda}_A \otimes \mathbf{1}_J$ is the orthogonal projection of $\ln \boldsymbol{p}$ onto $\mathcal{L}(C_A \otimes \mathbf{1}_J)$ and this is independent of the basis. The argument is the same for the other quantities.

A more usual notation for the log-linear model (see Fienberg (1989)) is to write $\ln p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$. These quantities relate to those defined above as follows: $u = \alpha_{11}/\sqrt{IJ}$, $u_{1(i)} = \phi_A \lambda_{Ai}/\sqrt{J}$, $u_{2(j)} = \phi_B \lambda_{Bj}/\sqrt{I}$ and $u_{12(ij)} = \phi_{AB}\lambda_{ABij}$.

3. Goodman's RC Model

Goodman (1979, 1981, 1985) introduced a family of log-multiplicative models which are especially suitable for contingency tables with ordered categories. The basic model of that family known as "the RC model" is given by

$$p_{ij} = \alpha_i \beta_j \exp[\phi \mu_i \nu_j].$$

Here μ_i and ν_j are parametric scores assigned to the categories of the table while α_i and β_j are nuisance parameters. The RC model implies the following appealing interaction form given by log odds

$$\log \frac{p_{ij} p_{ke}}{p_{ie} p_{kj}} = \phi(\mu_i - \mu_k)(\nu_j - \nu_e), \quad i \neq k, \quad j \neq e.$$

A variety of models can be obtained from the RC model. If for instance the scores are restricted to be equally spaced then the model states that the interaction between adjacent rows and adjacent columns is independent of the specific choice of such rows and columns. Further interpretation of the parameters of the RC model is given by Gilula (1986) and Gilula, Krieger and Ritov (1988).

The Goodman RC model is also a restricted form of (2) where λ_{AB} is no longer permitted to be a general unit vector in $\mathcal{L}\{\mathbf{c}_{ij} | i \neq 1, j \neq 1\}$ but is required to be of the form $\boldsymbol{\lambda}_{AB} = \boldsymbol{\mu} \otimes \boldsymbol{\nu}$ where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are unit vectors in $\mathcal{L}^{\perp}\{\mathbf{1}_I\}$ and $\mathcal{L}^{\perp}\{\mathbf{1}_J\}$, respectively. Thus the interaction term is assumed to have a specific mutliplicative structure. Hence the RC model is

$$\ln \boldsymbol{p} = \lambda^* \mathbf{c}_{1A} \otimes \mathbf{c}_{1B} + \phi_A \boldsymbol{\lambda}_A \otimes \mathbf{c}_{1B} + \phi_B \mathbf{c}_{1A} \otimes \boldsymbol{\lambda}_B + \phi_{AB} \boldsymbol{\mu} \otimes \boldsymbol{\nu}. \tag{7}$$

These parameters relate to those defined above as $\ln \alpha_i + \ln \beta_j = \lambda^*/\sqrt{IJ} + \phi_A \lambda_{Ai}/\sqrt{J} + \phi_B \lambda_{Bj}/\sqrt{I}$, $\phi = \phi_{AB}$ and μ_i, ν_j are the same. Note we can also write $\phi_{AB}^{1/2}\boldsymbol{\mu} = D_A\boldsymbol{\alpha}_A^*$, $\phi_{AB}^{1/2}\boldsymbol{\nu} = D_A\boldsymbol{\alpha}_B^*$ for $\boldsymbol{\alpha}_A^* \in \mathbf{R}^{I-1}$ and $\boldsymbol{\alpha}_B^* \in \mathbf{R}^{J-1}$ and in parallel with (1) we can write the model as

$$\ln \boldsymbol{p} = \alpha_{11}^* \boldsymbol{c}_{1A} \otimes \boldsymbol{c}_{1B} + (D_A\boldsymbol{\alpha}_A) \otimes \boldsymbol{c}_{1B} + \boldsymbol{c}_{1A} \otimes (D_B\boldsymbol{\alpha}_B) + D_A \otimes D_B\boldsymbol{\alpha}_B^* \otimes \boldsymbol{\alpha}_B^*. \quad (8)$$

An obvious adjustment to (6) leads to a formula for $\alpha_{11}^* = \lambda^*$. Further restrictions are sometimes placed on the μ and ν vectors; e.g. we might require $\mu_1 \leq \cdots \leq \mu_I$ and $\nu_1 \leq \cdots \leq \nu_J$ (see Ritov and Gilula (1991) for further discussion).

It should be noted here that originally the RC model was presented with no particular restrictions on $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Later, for identifiability and comparative purposes concerning the canonical correlation model, these parameters were restricted to have zero mean and unit variance with respect to the marginal distributions $\{p_{i\cdot}\}$ and $\{p_{\cdot j}\}$ respectively. Becker and Clogg (1989) correctly argued that making the restrictions depend upon marginal distributions induces some limitations when more than one contingency table is analyzed. Following their argument we choose here the two constraints of zero sum and unit length instead of the above expectations and variance constraints.

Suppose we specify a general $\boldsymbol{\lambda}_{AB} \in \mathcal{L}\{\boldsymbol{c}_{ij} | i \neq 1, j \neq 1\}$; then there are unique unit vectors $\boldsymbol{\mu} \in \mathcal{L}^{\perp}\{\mathbf{1}_I\}$ and $\boldsymbol{\nu} \in \mathcal{L}^{\perp}\{\mathbf{1}_J\}$ which minimize

$$\left\| \boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu} \right\|^2 \quad (9)$$

where $\|\cdot\|$ denotes Euclidean length, as is proven next.

Lemma 1. *For given $\boldsymbol{\lambda}_{AB}$ the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ specified above minimize* (9) *when they are respectively left and right singular vectors associated with the first singular value of the matrix*

$$\begin{matrix} \lambda_{11AB} & \cdots & \lambda_{1JAB} \\ \vdots & & \vdots \\ \lambda_{I1AB} & \cdots & \lambda_{IJAB} \end{matrix}$$

Proof. The proof follows straightforwardly from Eckart and Young (1936) and Householder and Young (1938).

Suppose we have specified ϕ_A, $\boldsymbol{\lambda}_A$, ϕ_B, $\boldsymbol{\lambda}_B$, ϕ_{AB} and $\boldsymbol{\lambda}_{AB}$, i.e. a particular log-linear model. Then Lemma 1 gives a method of fitting the Goodman RC model which is, in a certain sense, closest to this particular loglinear model. Further $\|\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}\|^2$ gives a measure of fit. Note that $0 \leq \|\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}\|^2 \leq 2$. As evident in the following section, these results are useful for the Bayesian analysis of the Goodman RC model. It should be noted, however, that, in the so-called classical ("frequentist") approach, the above mentioned measure of fit would be considered inferior because the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ obtained by the singular value decomposition are inefficient in the multinomial framework. Such singular value estimates can be used, however, in obtaining maximum likelihood estimates as is reported by Gilula and Haberman (1986, 1988).

A simple algorithm like the power method is available to compute $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. If we wish to find $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ which minimize (9) and also satisfy additional constraints beyond $\|\boldsymbol{\mu}\| = \|\boldsymbol{\nu}\| = 1$ and $1'\boldsymbol{\mu} = 1'\boldsymbol{\nu} = 0$ then a different optimizing algorithm is required. For example, if we add the order constraints then we must solve a nonlinear programming problem with a quadratic objective function and linear and quadratic constraints. For an alternative approach to a Bayesian analysis to tables with ordered variables see Agresti and Chuang (1989).

4. Bayesian Analysis

We begin by discussing the analysis of the full loglinear model. Let $F = (f_{ij})$ be the matrix of observed frequencies. The likelihood function for the full multinomial model is then

$$L(F|\boldsymbol{p}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{f_{ij}} \tag{10}$$

where $\boldsymbol{p}$ ranges in the $IJ - 1$ dimensional simplex. A Bayesian analysis then proceeds by placing a prior on $\boldsymbol{p}$, or some reparameterization, and then computing relevant posterior quantities. We discuss the choice of prior here.

The choice of parameterization materially affects the ease with which posterior calculations can be carried out. For example, a natural choice of a class of priors is given by the Dirichlet family which is conjugate. Hence with such a choice we can sample directly from the posterior and then do the necessary posterior computations for the ϕ and λ parameters.

Alternatively, analogous to random effect models, we can put a prior on the parameter α. The likelihood function in this parameterization is given by

$$L(F|\boldsymbol{\alpha}) \propto \exp\{\boldsymbol{\alpha}'C'\text{Vec}[F']\} = \exp\{\boldsymbol{\alpha}'\text{Vec}[(C_A'FC_B)']\} \tag{11}$$

where the Vec operator stacks the columns of a matrix. The constraint $\mathbf{1}'\boldsymbol{p} = 1$ specifies α_{11}, as in (6) while the remaining α_{ij} are completely free. The α_{i1}, $i \neq 1$, α_{1j}, $j \neq 1$ and α_{ij}, $i \neq 1$, $j \neq 1$ represents row, column and interaction contrasts respectively. Hence a reasonable choice for the prior is to take the α_{ij}, $(i,j) \neq (1,1)$ to be mutually statistically independent, $\alpha_{i1} \sim N(0, \alpha_A^2)$, $\alpha_{1j} \sim N(0, \sigma_B^2)$ and $\alpha_{ij} \sim N(0, \sigma_{AB}^2)$ as in a random effects model. The parameters σ_A, σ_B, σ_{AB} reflect our beliefs with respect to how much of the variability in the

$\ln p_{ij}$ is caused by row, column and interaction effects respectively. Note that with this choice of prior, $\phi_A \boldsymbol{\lambda}_A$, $\phi_B \boldsymbol{\lambda}_B$, and $\phi_{AB} \boldsymbol{\lambda}_{AB}$ are mutually statistically independent multivariate normals which are independent of the choice of C_A and C_B; e.g., $\phi_A \boldsymbol{\lambda}_A - N_I(0, \sigma_A^2 V)$ where $v_{ii} = 1 - 1/I$ and $v_{ij} = -1/I$ when $i \neq j$.

If we take σ_A^2, σ_B^2, and σ_{AB}^2 very large then this represents vague knowledge about column, row and interaction effects and we are effectively carrying out a likelihood analysis. For a diffuse analysis we consider a sequence of analyses with σ_A^2, σ_B^2, and σ_{AB}^2 becoming progressively larger until the posterior analysis stabilizes. Alternatively we could use data-dependent techniques for choosing the σ's. We mention one such approach, via chaining and the marginal likelihood, in Section 5. Also, we could treat σ_A^2, σ_B^2, and σ_{AB}^2 as hyperparameters, place a hyperprior on these quantities and integrate them out. This is the approach taken in Leonard (1975) in the context of the logistic model. An alternative approach to the analysis of a loglinear model is discussed in Epstein and Fienberg (1990).

With the parameterization of the log-linear model in terms of the α_{ij} and the above prior structure there is no simple algorithm available for sampling from the posterior for the log-linear model. An alternative approach for the computation is discussed in the following section. Note, however, that a diffuse analysis for the α_{ij} compels something like the above approach, for in choosing a Dirichlet for $\boldsymbol{p}$ it is not clear what choice should be made due to the change of variable.

If we accept the RC model as true in a given context, then a traditional approach puts a prior on $\boldsymbol{p}$, appropriately restricted, and then proceeds to compute various posterior quantities. The difficulty in determining the restrictions on $\boldsymbol{p}$ and thus an appropriate prior, however, makes this approach intractable. Perhaps more appropriately, as with the loglinear model, (voiding the problem of restrictions) is to put a prior on the parameters $(\alpha_A, \alpha_B, \alpha_A^*, \alpha_B^*)$. The likelihood function is then given by

$$\begin{aligned} &L(F|\boldsymbol{\alpha}_A, \boldsymbol{\alpha}_B, \boldsymbol{\alpha}_A^*, \boldsymbol{\alpha}_B^*) \\ &\propto \exp\left\{ \frac{N}{\sqrt{IJ}}\alpha_{11} + \frac{1}{\sqrt{J}}\boldsymbol{\alpha}_A' D_1 r + \frac{1}{\sqrt{I}}\boldsymbol{\alpha}_B' D_B c + \boldsymbol{\alpha}_A^{*\prime} D_A' F D_B \boldsymbol{\alpha}_B^* \right\} \end{aligned} \tag{12}$$

where $N = \mathbf{1}' F \mathbf{1}$, $r = F\mathbf{1}$ and $c = F'\mathbf{1}$.

This approach is more computationally difficult than with the loglinear model and this seems to be due to the likelihood function having a much more irregular shape.

A different approach is taken here for the Bayesian analysis of the RC model. With this approach computational difficulties are effectively subsumed within those for the loglinear model. Further, we avoid the assumption that the RC model is true and we provide a measure of how well the RC model fits when the full loglinear model is true.

For this, let $(\phi_A, \boldsymbol{\lambda}_A, \phi_B, \boldsymbol{\lambda}_B, \phi_{AB}, \boldsymbol{\lambda}_{AB})$ be the parameter for the full log-linear model and let $E[\cdot|F]$ denote the expectation operator with respect to the posterior distribution of the full model. Let $(\phi_A, \boldsymbol{\lambda}_A, \phi_B, \boldsymbol{\lambda}_B, \phi_{AB}, \boldsymbol{\mu}, \boldsymbol{\nu})$ denote the parameter of the unique RC model which is closest to the above log-linear model in the sense described in Section 2. Then the posterior distribution on $(\phi_A, \boldsymbol{\lambda}_A, \phi_B, \boldsymbol{\lambda}_B, \phi_{AB}, \boldsymbol{\lambda}_{AB})$ induces a marginal posterior distribution on $(\phi_A, \boldsymbol{\lambda}_A, \phi_B, \boldsymbol{\lambda}_B, \phi_{AB}, \boldsymbol{\mu}, \boldsymbol{\nu})$ and we can use this to make inferences about the RC model. Further, we can look at

$$E\Big[||\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}||^2 |F\Big] \tag{13}$$

to assess how well the RC model fits the data F. A value near 2 indicates a poor fit while a value near 0 indicates a good fit. Of course, we also have the full marginal posterior distribution of $||\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}||^2$ for this kind of inferences as well.

5. Computations

Our approach to computing integrals is via Monte Carlo and in particular, importance sampling and adaptive importance sampling. A Monte Carlo approach seems necessary for the analysis of the RC model which we presented in Section 3. Adaptive importance sampling has been discussed by many authors; in particular see Smith et al. (1987), Evans (1991a,b), Oh and Berger (1989) and for an implementation in the context of contingency tables see Evans, Gilula and Guttman (1989).

For the analysis of the loglinear model we must evaluate ratios of the form

$$\frac{\int g(\boldsymbol{\alpha})L(F|\boldsymbol{\alpha})h(\boldsymbol{\alpha})d\boldsymbol{\alpha}}{\int L(F|\boldsymbol{\alpha})h(\boldsymbol{\alpha})d\boldsymbol{\alpha}}, \tag{14}$$

where the integration is over R^{IJ-1}, where h is the prior on $\boldsymbol{\alpha}$ for specified values of σ_A^2, σ_B^2 and σ_{AB}^2, and g is some real-valued function; e.g. $g(\boldsymbol{\alpha}) = \alpha_{ij}^2$. Hence, we must approximate the integrals in the numerator and the denominator.

Asymptotically, under quite general conditions, the posterior distribution can be well approximated by a multivariate normal distribution with mean at the mode $\hat{\boldsymbol{\mu}}$ of the posterior and with variance matrix $\hat{\Sigma}$ equal to the negative of the inverse of the Hessian matrix of the log-posterior, evaluated at $\hat{\boldsymbol{\mu}}$. For a diffuse analysis these quantities are effectively the maximum likelihood estimates. With this approximation we can generate $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from this normal distribution and estimate $E[g|F]$ by $\sum_{i=1}^n g(\boldsymbol{X}_i)/n.$. In many contexts, the tails of the posterior are longer than normal tails and thus these Monte Carlo estimates will have infinite variance. Hence a conservative approach is to use a multivariate Student density function with mean given by $\hat{\mu}$, variance matrix given by $\hat{\Sigma}$ and degrees of freedom given by τ, for the Monte Carlo estimates in place of the multivariate normal. It is difficult to determine a precise value for τ but it seems better to err on the side of conservatism. More generally, the posterior distribution may have an extremely non-normal shape (see Naylor and Smith (1988) and Smith et al. (1987)).

Further, for small data sets we can expect the posterior mean vector and variance matrix to differ from $\hat{\mu}$ and $\hat{\Sigma}$, respectively. Hence, it makes sense to estimate these quantities via importance sampling, replace our initial estimates and to continue this procedure for several steps. This is called adaptive importance sampling and is more extensively discussed in the references given above. For example, however, Oh and Berger (1989) show that this method improves the efficiency of the computation over straight importance sampling.

The above presupposes that we have chosen values for σ_A^2, σ_B^2 and σ_{AB}^2. For a diffuse analysis we want to choose these quantities to be large. To assess whether or not we have chosen these values large enough we must do the posterior analysis for several choices of these parameters and observe how the analysis changes. If appropriate choices have been made then choosing even larger values should produce negligible changes in the estimates of the posterior quantities of interest; e.g. the means and variances of the α_{ij}. In other contexts we may wish to choose the σ parameters to depend on the data. For example the denominator in (14) can be viewed as a marginal likelihood for $(\sigma_A^2, \sigma_B^2, \sigma_{AB}^2)$, and it is reasonable then to select the values of these parameters to maximize this quantity. We do not pursue this approach further here, as our interest is in a diffuse analysis, but note that the methods discussed in Evans (1991a) may be appropriate in this context.

6. Examples

Example 1. We consider the implementation of the analysis with a specific data set taken from Goodman (1985, Table 2). The purpose of the analysis is to examine the relationship between a person's mental health status and the parents' socioeconomic status. The data are recorded in Table 1. This is a 4 × 6 table and hence the integrals required for a Bayesian analysis are 23 dimensional.

We carry out a diffuse analysis with $\sigma_A^2 = \sigma_B^2 = \sigma_{AB}^2 = 100$. In columns 2 and 3 of Table 2 we have recorded the values for the means and standard deviations for the starting importance sampler. The starting covariances are not given for reasons of space. As the data set is based on a sample of $N = 1660$ we might expect the normal approximation to hold quite well in this example and, indeed, it does. Hence, we chose $\tau = 20$ and used straight importance sampling with no adaptation. The final two columns in Table 2 give the estimated posterior means and standard deviations for the α_{ij} based on a Monte Carlo sample of $n = 10,000$. Note that there is very little change in these values. To assess the accuracy of these estimates we generated a further sample of 50,000 and compared these estimates. Based on this we are satisfied that the posterior means and standard deviations have been computed to 2 decimal places of accuracy. The estimates of Table 2 required about 30 seconds of CPU time. Also using a multivariate normal importance sampler gave similar results. To check that this was indeed a diffuse analysis we did the computations with $\sigma_A^2 = \sigma_B^2 = \sigma_{AB}^2 = 1000$ and again no significant changes in the results were observed.

The posterior expectation of ϕ_{AB} was computed to be .133 and the posterior standard deviation .359. Table 3 records some posterior quantiles for ϕ_{AB}. Hence we see that the posterior distribution is skewed to the right and there is ample evidence that $\phi_{AB} \neq 0$.

The posterior expectation and standard deviation of $\|\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}\|^2$ were computed to be .137 and .153 respectively. Table 4 records some posterior quantiles for this quantity. Note that the posterior distribution is skewed to the right in $[0, 2]$ but reasonably concentrated about 0. Hence there is evidence that the RC model is providing a reasonable fit for this data. This is compatible with Goodman's results (Goodman (1985)). Note that Monte Carlo seems to be a necessity for the computation of this distribution.

Given that the RC model provides a good fit we record the posterior expectations of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in Table 5. Note that, to a remarkable degree, these scores are ordered in a way which agrees with the ordering of the categories of the variables. Hence the RC model does well in explaining the lack of independence in this context. The computations in Tables 3, 4 and 5 are based on a further sample of size 10,000.

Example 2. This example comes from Goodman (1981). Here, 135 subjects are cross-classified by Periodontal Condition (row variable) and by Calcium intake level (column variable). These data are given in Table 6 where A denotes the best condition and 1 is the lowest calcium intake.

The integrals in this problem are all 15-dimensional. Because of the somewhat smaller sample size of $N = 135$ we chose a 15-dimensional Student distribution with $\tau = 5$ as the basic kernel and used adaptive importance sampling, adapting to the mean and variance matrix of the posterior. Thus we started the importance sampler using a 15-dimensional Student with 5 degrees of freedom, mean vector given by $\hat{\boldsymbol{\mu}}$ and with variance matrix given by $\hat{\Sigma}$. Then we updated these estimates after each Monte Carlo sample of size 1000. For example, if N_i and D_i represent the estimates of the numerator and denominator of (14) at the ith step of this process, where $g(\boldsymbol{\alpha}) = \alpha_{12}$ and n_i, d_i represent the estimates of these quantities based on the ith sample of 1000, then $N_{i+1} = (iN_i + n_i)/(i+1)$, and similarly for D_{i+1}. Hence, the ith sample was generated from a multivariate Student whose first coordinate has mean N_i/D_i, and the $(i+1)$-st sample was generated from a multivariate Student whose first coordinate has mean N_{i+1}/D_{i+1}. We update the means, variances and covariances of all the coordinates of the importance sampling distribution using this technique. Our estimates of the posterior mean and variance matrix are based on 10 steps of this process, or 10,000 generated values.

The starting values and the final estimates are recorded in the first 4 columns of Table 7. Comparing the results in columns 3 and 4 with a simulation of size 50,000 we estimate that the largest relative error in the means is about 5% and the average relative error in the means is about 2%. The corresponding quantities for the standard deviations are about 8% and 3% respectively. The changes from strating to final estimates are not dramatic but we note that all the posterior standard deviations are larger than the starting values indicating that a multivariate *normal* importance sampler might not effectively cover the region where posterior probability lies. Also, note that the changes in the means pertaining to columns 1 and 3 indicates that the posterior is skewed as the starting value is the mode.

In the final two columns of Table 7 we give estimates of the posterior means and standard deviations which were computed using a sample of 10,000 from a multivariate normal with mean $\hat{\mu}$ and variance $\hat{\Sigma}$. These do not differ substantially from our previous estimates, but an error analysis indicates a maximum relative error of about 6% and average relative error of about 3% for the means with 12% and 4% being the relevant quantities for the standard deviations. Hence, the adaptive importance sampling algorithm is more accurate but not greatly so in this case. In general, however, we feel it is better to be conservative and use an importance sampler in which we feel confident that its distribution effectively covers a region containing the bulk of the posterior probability and also to use adaptive techniques. Although adaptation has not produced large improvements in computational efficiency in this problem it has done so in other contexts. Further adaptation adds only a few seconds to our computation times so it is relatively inexpensive.

The posterior expectation and standard deviation of ϕ_{AB} were computed to be 7.270 and 7.540 respectively. Hence, it would seem that periodontal condition and calcium intake are not independent; but there is a large degree of uncertainty in this inference. If we assume they are not independent then it makes sense to fit a RC model. The posterior expectation and standard deviation of $||\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}||^2$ are .056 and .093 respectively. Hence, it seems clear that an RC model is providing a good fit. The posterior expectations of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are given in Table 8. From this we see the scores for periodontal are monotonic but not for calcium intake, as the last two categories reverse their order. It can be shown (as in Goodman (1981)) that the RC model also fits the collapsed table obtained from the original table by combining the last two columns (where monotonicity of order is violated). In the collapsed table both set of scores come out monotone.

Acknowledgement

The authors would like to thank a referee and the Editor for insightful comments leading to a significant improvement of a previous version.

Table 1. Cross-classification of subjects by their mental health status and by the socioeconomic status of their parents

	Parents' socioeconomic status					
Mental health status	A	B	C	D	E	F
Well	64	57	57	72	36	21
Mild symptom formation	94	94	105	141	97	71
Moderate symptom formation	58	54	65	77	54	54
Impaired	46	40	60	94	78	71

Table 2. Posterior means and standard deviations for the α_{ij} in Example 1

Variable	Mean	Std. dev.	Mean	Std. dev.
	(Start)	(Start)	(Finish)	(Finish)
$(1,2)$	−0.116	0.131	−0.115	0.130
$(1,3)$	0.220	0.125	0.222	0.124
$(1,4)$	0.652	0.112	0.655	0.112
$(1,5)$	−0.214	0.130	−0.217	0.129
$(1,6)$	−0.606	0.149	−0.612	0.149
$(2,1)$	1.264	0.130	1.272	0.130
$(2,2)$	0.059	0.118	0.058	0.118
$(2,3)$	0.097	0.116	0.097	0.116
$(2,4)$	0.106	0.106	0.110	0.106
$(2,5)$	0.284	0.133	0.289	0.133
$(2,6)$	0.379	0.167	0.384	0.169
$(3,1)$	−0.264	0.130	−0.265	0.130
$(3,2)$	−0.008	0.128	−0.007	0.130
$(3,3)$	0.082	0.124	0.081	0.124
$(3,4)$	−0.009	0.114	−0.008	0.115
$(3,5)$	0.124	0.134	0.126	0.134
$(3,6)$	0.418	0.147	0.423	0.148
$(4,1)$	−0.106	0.130	−0.108	0.129
$(4,2)$	−0.047	0.146	−0.047	0.146
$(4,3)$	0.190	0.133	0.191	0.135
$(4,4)$	0.296	0.114	0.300	0.112
$(4,5)$	0.452	0.122	0.458	0.125
$(4,6)$	0.519	0.131	0.525	0.131

Table 3. Posterior quantiles for ϕ_{AB} in Example 1

Probability	Quantile
.010	.639
.025	.713
.100	.894
.500	1.293
.900	1.797
.975	2.133
.990	2.310

Table 4. Posterior quantiles for $\|\boldsymbol{\lambda}_{AB} - \boldsymbol{\mu} \otimes \boldsymbol{\nu}\|^2$ in Example 1

Probability	Quantile
.010	.029
.025	.038
.010	.063
.500	.126
.900	.225
.975	.303
.990	.339

Table 5. Posterior expectations of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in Example 1

$\mu_1 = -.413$	$\nu_1 = -.248$
$\mu_2 = -.017$	$\nu_2 = -.255$
$\mu_3 = .059$	$\nu_3 = -.091$
$\mu_4 = .370$	$\nu_4 = -.009$
	$\nu_5 = .202$
	$\nu_6 = .401$

Table 6. Periodontal condition by calcium intake

	Calcium intake			
Periodontal condition	1	2	3	4
A	5	3	10	11
B	4	5	8	6
C	26	11	3	6
D	23	11	1	2

Table 7. Posterior means and standard deviations of the α_{ij} in Example 2

Variable	Mean (Start)	Std. dev. (Start)	Mean (Finish)	Std. dev. (Finish)	Mean (Normal)	Std. dev. (Normal)
$(1,2)$	−0.667	0.395	−0.722	0.444	−.712	.408
$(1,3)$	−1.211	0.558	−1.388	0.626	−1.374	.590
$(1,4)$	−0.339	0.481	−0.348	0.514	−.347	.523
$(2,1)$	−0.191	0.428	−0.180	0.458	−.200	.446
$(2,2)$	0.367	0.496	0.434	0.522	.411	.523
$(2,3)$	−0.212	0.396	−0.224	0.419	−.222	.421
$(2,4)$	−0.384	0.383	−0.418	0.384	−.407	.410
$(3,1)$	0.575	0.407	0.596	0.426	.590	.432
$(3,2)$	−0.414	0.354	−0.415	0.375	−.403	.358
$(3,3)$	−1.663	0.464	−1.777	0.501	−1.801	.508
$(3,4)$	−0.618	0.398	−0.656	0.415	−.644	.420
$(4,1)$	−0.598	0.592	−0.802	0.682	−.773	.622
$(4,2)$	−0.217	0.311	−0.206	0.326	−.209	.327
$(4,3)$	−1.909	0.751	−2.217	0.864	−2.177	.823
$(4,4)$	−0.955	0.625	−1.043	0.656	−1.028	.680

Table 8. Posterior expectations of μ and ν in Example 2

$\mu_1 = 0.2128452$	$\nu_1 = -0.230028$
$\mu_2 = 0.1572506$	$\nu_2 = -0.161415$
$\mu_3 = -0.102727$	$\nu_3 = 0.2566262$
$\mu_4 = -0.267369$	$\nu_4 = 0.1348174$

References

Agresti, A. and Chuang, C. (1989). Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. *Comput. Statist. Data Anal.* 7, 245–258.

Becker, M. P. and Clogg, C. C. (1989). Analysis of sets of two-way contingency tables using association models. *J. Amer. Statist. Assoc.* **84**, 142–151.

Eckart, C. and Young, G. (1936). The approximation of a matrix by one of lower rank. *Psychometrika* **1**, 211–218.

Epstein, L. D. and Fienberg, S. E. (1990). Bayesian inference for loglinear models in contingency tables. Paper presented at Data Analysis and Statistical Foundations Conference, University of Toronto, May 31 - June 1, 1990.

Evans, M. (1991a). Chaining via annealing. *Ann. Statist.* **19**, 382–393.

Evans, M. (1991b). Adaptive importance sampling and chaining. *Contemporary Mathematics*, 115, Statistical Multiple Integration (Edited by N. Flournoy and R. Tsutuakawa), 137–143, Amer. Math. Soc.

Evans, M. J., Gilula, Z. and Guttman, I. (1989). Latent class analysis of two-way contingency tables by Bayesian methods. *Biometrika* **76**, 557-563.

Fienberg, S. E. (1989). *The Analysis of Cross-Classified Categorical Data*, 2nd edition. The MIT Press.

Gilula, Z. (1986). Grouping and association in contingency tables: An exploratory canonical correlation approach. *J. Amer. Statist. Assoc.* **81**, 773-779.

Gilula, Z. and Haberman, S. J. (1986). Canonical analysis of contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.* **81**, 780-788.

Gilula, Z. and Haberman, S. J. (1988). The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist. Assoc.* **83**, 760-771.

Gilula, Z., Krieger, A. M. and Ritov, Y. (1988). Ordinal association in contingency tables: Some interpretive aspects. *J. Amer. Statist. Assoc.* **83**, 540-545.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74**, 537-552.

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76**, 320-334.

Goodman, L. A. (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. *Ann. Statist.* **13**, 10-69.

Haberman, S. J. (1974a). *The Analysis of Frequency Data.* University of Chicago Press, Chicago.

Haberman, S. J. (1974b). Log-linear models for frequency tables with ordered classifications. *Biometrics* **30**, 589-600.

Householder, A. S. and Young, G. (1938). Matrix approximation and latent roots. *Amer. Math. Monthly* **45**, 165-167.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. Ser.B* **37**, 23-37.

Leonard, T., Hsu, J. S. J. and Tsui, K.-W. (1989). Bayesian marginal inference. *J. Amer. Statist. Assoc.* **84**, 1051-1058.

Naylor, J. C. and Smith, A. F. M. (1988). Econometric illustrations of novel numerical integration strategies for Bayesian inference. *J. Econom.* **38**, 103-125.

Oh, M. S. and Berger, J. O. (1989). Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simul.* (To appear)

Ritov, Y. and Gilula, Z. (1991). The order-restricted RC model for ordered contingency tables: Estimation and testing for fit. *Ann. Statist.* **19**, 2090-2101.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987). Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician* **36**, 75-82.

Department of Statistics, University of Toronto, Toronto, On M5S 1A1, Canada.
Department of Statistics, Hebrew University, Jerusalem 91905, Israel.

(Received August 1990; accepted December 1992)