Bayesian inference for square contingency tables

Jonathan J. Forster

SUMMARY

Inference for multivariate categorical data often proceeds by selecting a log-linear model from a set of competing models or, in a Bayesian approach, by averaging inferences over the set, weighted by posterior probabilities. In this paper, we use permutation invariance as a criterion for constructing a set of models for this purpose, for the common situation when the data form a ‘square’ contingency table, representing a cross-classification by two categorical variables with identical categories.

We consider log-linear models which are invariant under certain groups of permutations of the cells of a multiway contingency table. We present permutation invariant log-linear models for a number of different contingency table structures and show how to construct invariant prior distributions for the model parameters.

Keywords: Bayesian analysis; Contingency table; Group representation; Permutation group; Prior distribution; Square table

1 Introduction

A natural way of representing the distribution of categorical data is via a vector of probabilities $p$, with each component representing the probability of observing a particular category. Usually, observations are made on a set $C$ of categorical variables, and the multivariate data can be represented as the cell counts of a $|C|$-way contingency table. Using the notation of Darroch, Lauritzen and Speed (1980), the table is the set $I = \prod_{c \in C} I_c$ where $I_c$ is the set of levels of factor $c$. An individual cell is denoted by $i = (i_c, c \in C)$, and $p = (p(i), i \in I)$. The total number of cells in the table is $m = |I| = \prod_{c \in C} |I_c|$ and $p$ lies in the $(m-1)$-dimensional simplex

$$S_{m-1} = \left\{ \mathbb{R}^m : \sum_{i=1}^m p_i = 1; \ p_i \geq 0, \ i = 1, \ldots, m \right\}.$$ 

For a $k$-way table, we will assume the notation $C = \{1, \ldots, k\}$ and $|I_i| = r_i$ for $i = 1, \ldots, k$ so that $m = \prod_{i=1}^k r_i$.

---

1Department of Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK email: jjf@maths.soton.ac.uk
It is often assumed that individuals are classified independently, and therefore, if a total number $N$ of individuals are observed then the distribution of the corresponding ‘cell counts’ $n = (n(i), i \in I)$ is multinomial. It is sometimes convenient to treat $N$ as coming from a Poisson distribution, in which case the $n(i)$ are also Poisson, and independent. The distribution of $n$ is then represented by the corresponding vector of cell means $\mu = (\mu(i), i \in I)$. However, these two approaches are essentially equivalent, and the real interest is focussed on models for $p$ or $\mu$. In what follows, we shall consider models for $p$, noting any circumstances where models for $\mu$ are substantively different.

Because of the awkward constraints imposed by $p$ lying in $S_{m-1}$, we consider the alternative parameterisation in terms of the centred log ratio $\theta_i = \log(p_i) - \frac{1}{m} \log(p_1 p_2 \cdots p_m)$ $i = 1, \ldots, m$ (Aitchison, 1986). Clearly, $\theta \in 1^\perp_m = \{\mathbb{R}^m : 1^T_m \theta = 0\}$, a $(m-1)$-dimensional vector space, which proves far easier to work with.

Association between categorical variables may be investigated by examining the structure of a contingency table using log-linear models. A log-linear model for $p$, expressed in terms of $\theta$, specifies that $\theta = X\beta$, where $X$ is a model matrix of known coefficients with columns orthogonal to $1_m$ and $\beta$ is a vector of model parameters. Alternatively this model may be expressed in ‘constraint form’, $A\theta = 0$ for an appropriate constraint matrix $A$. The row space of $A$ is the orthogonal complement to the column space of $X$. The column space of $X$ may be thought of as a subspace of ‘realisation space’, the set of all possible values taken by $\theta$, whereas the row space of $A$ is a subspace of the dual vector space, ‘log-contrast space’, the set of all possible log-contrasts of $p$. Strictly speaking, not all log-linear models for $p$ can be expressed in terms of log-contrasts, but those that cannot are rarely of interest.

Log-linear models can therefore be thought of as vector subspaces of $1^\perp_m$, with the interpretation of a subspace either as a space where the realisations are constrained to lie, or as a space of log-contrasts, each of which must be satisfied by any realisations. There are clearly therefore an infinite number of possible log-linear models which may be considered in any situation. This set may be reduced by considering only those models which are invariant to certain permutations of the indices in $I$. This is desirable in any situations where the prior belief about the cell probabilities is unaltered when the indices in $I$ are permuted in certain ways. Clearly, any Bayesian analysis requires an explicit statement of prior belief, so identification of such a model class is essential. However, any approach which seeks to identify plausible models needs to be aware of the dangers of providing a quantitative assessment of the appropriateness of any model which has been suggested by observing the data, so a reference set of a priori plausible models is still desirable. Even in cases where
the permutations are ‘too restrictive’ and overstate the degree of prior uncertainty, the set of models obtained will still provide a useful reference.

2 Models

Often, two or more categories will have the same levels. Such data is common in social research where couples or parent/child pairs may be cross-classified according to an attribute recorded for each of the pair. Although this data is not restricted to two-way tables, it is most commonly in that form and we shall focus on the two-way ‘square’ contingency table.

Consider a \( k \)-way table, where now each of the \( k \) variables takes the same \( r \) levels. It no longer seems reasonable to insist that models should be invariant under an arbitrary combination of permutations of the levels of each of the variables. A more reasonable restriction is to require that models are invariant to the same simultaneous permutation of the \( r \) equivalent levels of each variable. Furthermore, it may now be considered that, as the variables have a similar interpretation, then invariance to permuting the variables themselves should be required. Hence, in this situation, the required permutation group is the direct product \( G = S_r \times S_k \), as any permutation of factors commutes with permutation of factor labels. For a two-way table, \( G = S_r \times C_2 \). Then the group actions are the simultaneous permutation of row and column category levels, represented by the matrix \( P_{g_i} \otimes P_{g_i} \), where \( P_{g_i} \) is the \( r \times r \) permutation matrix representing permutation \( g_i \); the reflection of the table about its main diagonal, represented by the matrix \( P_h \) which leaves diagonal cells unchanged and transposes all symmetrically opposite pairs of off-diagonal cells; together with the corresponding products.

It is immediately clear that there are two orbits of this group. Diagonal cells always remain on the diagonal, and off-diagonal cells always remain off-diagonal. Initially, we will consider these two groups of cells separately. This is consistent with many standard approaches to modelling this kind of data. Indeed, in some examples, such as paired comparisons data, the diagonal cells are structural zeros, and are therefore excluded from analysis. Suppose that the diagonal cells are denoted \( 1, \ldots, r \), and the off-diagonal cells are denoted \( r+1, \ldots, r^2 \), then \( \theta^d \in 1_r^+ \) with components

\[
\theta^d_i = \theta_i - \frac{1}{r} \sum_{j=1}^r \theta_j \quad i = 1, \ldots, r
\]

is the centred logratio for the conditional diagonal cell probabilities, and \( \theta^d \in 1_{r(r-1)}^+ \) with components

\[
\theta^d_i = \theta_i - \frac{1}{r(r-1)} \sum_{j=r+1}^{r^2} \theta_j \quad i = 1, \ldots, r(r-1)
\]
is the centred logratio for the conditional diagonal cell probabilities.

Therefore

\[ 1_m^\perp = V_1 \oplus 1_r^\perp \oplus 1_{r(r-1)}^\perp \]  

(1)

is a decomposition of \(1_m^\perp\) into three mutually orthogonal \(G\)-invariant subspaces, where \(V_1\) is the orthogonal complement of \(1_r^\perp \oplus 1_{r(r-1)}^\perp\) in \(1_m^\perp\), which is spanned by the vector \(v\) taking values \((r - 1)/r\) for the diagonal cells and \(-1/r\) for the off-diagonal cells. We consider each of these orthogonal \(G\)-invariant subspaces in turn.

The subspace \(V_1\) is 1-dimensional, and corresponds to the trivial representation of \(S_r \times C_2\), denoted \([r] \otimes [2]\). In the orthogonal decomposition \(\theta = \theta v + \theta d + \theta \bar{d}\), the component \(\theta\) is given by

\[ \theta = \frac{1}{r} \sum_{j=1}^{r} \theta_j - \frac{1}{r(r-1)} \sum_{j=r+1}^{r^2} \theta_j. \]  

(2)

The \(G\)-invariant model \(\theta \in V_1\) implies homogeneity of diagonal cell probabilities and homogeneity of off-diagonal cell probabilities, with the single parameter \(\theta\) controlling the total probability on the diagonal.

The action of \(G\) on the diagonal cells and \(\theta d \in 1_r^\perp\) is isomorphic to \(S_r\), as the reflection leaves these cells unchanged. As stated in Section 3.1, \(1_r^\perp\) is an irreducible \(S_r\)-invariant subspace, and there are just two \(G\)-invariant log-linear models for the (conditional) diagonal cell probabilities, the saturated model, and the model where they are all constrained to be the same. The irreducible invariant subspace \(1_{r}^\perp\), which we shall call \(V_{21}\) corresponds to the representation of \(G\) denoted \([r-11] \otimes [2]\).

Now consider the off-diagonal cell probabilities. We define the \(r^2 \times r^2\) relationship matrices \(E\) (opposites), \(F\) (neighbours) and \(G\) (opposite neighbours) as follows, denoting off-diagonal cells by both row and column subscripts.

\[
E_{ijkl} = \begin{cases} 
1 & \text{if } i = l \text{ and } j = k \\
0 & \text{otherwise}
\end{cases}
\]

\[
F_{ijkl} = \begin{cases} 
1 & \text{if } i = k \text{ and } j \neq l \\
0 & \text{otherwise}
\end{cases}
\]

\[
G_{ijkl} = \begin{cases} 
1 & \text{if } j = l \text{ and } i \neq k \\
0 & \text{otherwise}
\end{cases}
\]

All elements of \(E\), \(F\) and \(G\) corresponding to cells on the diagonal of the table are zero. Each of these matrices represents relationships which are preserved under the action of \(G\), and so each of \(E\), \(F\) and \(G\) commute with every \(P_G\). Together with \(I\) and \(J\), these matrices form a basis for a \(G\)-invariant relationship algebra (James, 1957).
There are four $G$-invariant irreducible subspaces of $1_{\frac{r}{(r-1)}}^\perp$, the parameter space for the off-diagonal cell probabilities. They are most easily described by their projections

\[
Q_{22} = \frac{1}{2(r-2)}(2I + 2E + F + G) - \frac{2}{r(r-2)}J
\]

\[
Q_3 = \frac{1}{2r}(2I - 2E + F - G)
\]

\[
Q_4 = \frac{1}{2}(I + E) - \frac{1}{r(r-1)}J - Q_2
\]

\[
Q_5 = \frac{1}{2}(I - E) - Q_3
\]

These matrices are all clearly symmetric and idempotent, and can be derived using Table 5 of Brien et al (1988). Invariance follows from the fact that $Q_{22}, Q_3, Q_4, Q_5$ are members of the invariant relationship algebra spanned by $I, J, E, F$ and $G$ and therefore commute with every $P_g$. Irreducibility can be determined by considering group characters, as mentioned in Section 2. We denote the four invariant irreducible log-linear models for off-diagonal cell means by subspaces $V_{22}, V_3, V_4, V_5$, which are of dimension $(r-1), \frac{1}{2}(r-3)$ and $\frac{1}{2}(r-1)(r-2)$ respectively, corresponding to $Q_{22}, Q_3, Q_4, Q_5$. These invariant subspaces are also inequivalent, corresponding to representations denoted $[r-1] \otimes [2], [r-1] \otimes [1 1], [r-2] \otimes [2]$ and $[r-2] \otimes [1 1]$ respectively. Therefore there are at most 16 $G$-invariant log-linear models for off-diagonal cell probabilities.

Some of the invariant log-linear models are familiar and may easily be expressed in terms of constraints on the parameters of a saturated log-linear model for off-diagonal $p$ written as

\[
\log p_{ij} = \begin{cases} 
\mu + \alpha_i + \alpha_j + \beta_i - \beta_j + \gamma_{ij} + \gamma_{ji} + \lambda_{ij} - \lambda_{ji} & i < j \\
\mu + \alpha_i + \alpha_j + \beta_i - \beta_j + \gamma_{ij} + \gamma_{ji} - \lambda_{ij} + \lambda_{ji} & i > j 
\end{cases}
\] (3)

where the subscripts correspond to row and column level. Here $\mu$ is a normalising constant, and the other parameters are expressed in a suitably ‘symmetric’ form. The invariant subspaces $V_{22}, \ldots, V_5$ can be thought of as parameter spaces for $\alpha, \beta, \gamma$ and $\lambda$ respectively. Familiar models include:

$V_{22} \oplus V_4$ \hspace{1cm} **Symmetry**

\[
\lambda_{ij} = 0 \quad i < j \\
\beta_i = 0 \quad i = 1, \ldots, r,
\]

$V_{22} \oplus V_4 \oplus V_5$ \hspace{1cm} **Quasi-Symmetry**

\[
\lambda_{ij} = 0 \quad i < j,
\]
\[ V_{22} \oplus V_3 \quad \text{Quasi-Independence} \quad (\text{of off-diagonal cells}) \]
\[
\begin{align*}
\lambda_{ij} &= 0 \quad i < j \\
\gamma_{ij} &= 0 \quad i < j,
\end{align*}
\]

\[ V_{22} \quad \text{Quasi-Independence + Marginal Homogeneity} \]
\[
\begin{align*}
\lambda_{ij} &= 0 \quad i < j \\
\gamma_{ij} &= 0 \quad i < j,
\end{align*}
\]

together with the saturated model \((1_{r(r-1)}^1 = V_{22} \oplus V_3 \oplus V_4 \oplus V_5)\) and the null model (equiprobability; 0). Here the models are expressed in ‘realisation space’. The required subspaces for the ‘log-contrast’ representation are the corresponding orthogonal complements (e.g. \(V_{22} \oplus V_4\) for Symmetry). For further details of these models, see Agresti (1990), Becker (1990) and von Eye and Spiel (1996).

Now consider the whole contingency table, and its parameter space \(1_m^{\perp}\). The decomposition \((??)\), together with the decomposition of \(1_{r(r-1)}^1\) described above gives the \(G\)-invariant irreducible decomposition \(1_m^{\perp} = V_1 \oplus V_21 \oplus V_22 \oplus V_3 \oplus V_4 \oplus V_5\). Each of the off-diagonal models described above may be considered as models for the whole table, by including \(V_1 \oplus V_21\).

Although the decomposition \(1_m^{\perp} = V_1 \oplus V_21 \oplus V_22 \oplus V_3 \oplus V_4 \oplus V_5\) is irreducible, it is not unique, as \(V_21\) and \(V_22\) are both \((r-1)\)-dimensional subspaces corresponding to the same irreducible representation \([r-1 1] \otimes [2]\). Wherever there are a pair of equivalent invariant subspaces, such as \(V_21\) and \(V_22\) then for any \(\phi \in (-\pi/2, \pi/2]\), a further equivalent subspace is \(V_2(\phi)\), spanned by vectors of the form \(v_{21} \cos \phi + v_{22} \sin \phi\) where \(v_{21}\) and \(v_{22}\) are basis elements, in \(V_21\) and \(V_22\) respectively, which are equivalent under the representation concerned. In order to completely specify \(\phi\), we assume, without loss of generality, that the two bases are orthonormal in \(1_m^\perp\). Therefore \(V_21 = V_2(0)\) and \(V_22 = V_2(\pi/2)\), and the contribution \(V_21 \oplus V_22\) in the decomposition of \(1_m^{\perp}\) may be replaced by \(V_2(\phi_1) \oplus V_2(\phi_2)\) for any distinct \(\phi_1, \phi_2 \in (-\pi/2, \pi/2]\).

For example, the model of independence of the factors in the two-way table is represented by \(\theta \in V_3 \oplus V_2(\arccos \sqrt{2/r})\). The subspace \(V_2(\arccos \sqrt{2/r})\) may be thought of as the parameter space for \(\alpha\) in the expression of the independence model

\[
\log p_{ij} = \mu + \alpha_i + \alpha_j + \beta_i - \beta_j. \quad (4)
\]
Although the model of independence is rarely of interest for a square table, one model which is of interest, and which involves $V_2(\arccos \sqrt{2/r})$ is the diagonal parameter model (Agresti 1988; Tanner and Young, 1985)

$$\log p_{ij} = \begin{cases} 
\mu + \alpha_i + \alpha_j + \beta_i - \beta_j + \gamma & i = j \\
\mu + \alpha_i + \alpha_j + \beta_i - \beta_j & i \neq j 
\end{cases} \tag{5}$$

The invariant subspace corresponding to this model is $V_1 \oplus V_3 \oplus V_2(\arccos \sqrt{2/r})$. Here, $V_3$ may be thought of as the parameter space for $\gamma$.

Another situation where such considerations arise is for Poisson log-linear models which correspond to subspaces of $\mathbb{R}^m = 1_m \perp 1_m \oplus 1_m$. The extra irreducible subspace of $\mathbb{R}^m$, namely $1_m$ corresponds to the same irreducible representation $[r] \otimes [2]$ as $V_1$. As these two subspaces are both one-dimensional, it is easier to describe the implications of this. Any linear combination of $1$ and $v$ will also be $G$-invariant, so any two linearly independent such combinations may be considered as the bases of two invariant subspaces. Perhaps the most immediately obvious such alternative subspaces would be spanned by $1/r + v$ and $(r - 1)/r - v$, which indicate the diagonal and off-diagonal cells respectively.

### 3 Invariant Prior Distributions

The consideration of invariance is perhaps most critical when a Bayesian analysis is intended. Then, it is important that prior distributions are constructed so as to respect invariance to any transformation under which prior belief remains unchanged. In many categorical data analyses, prior belief may not change when the labels of the factors are permuted in certain ways. The prior distribution for the cell probabilities should be constructed in a way which respects such considerations. The resulting prior distributions tend to reflect prior belief about structural models, rather than sizes of individual cell probabilities.

As previously, we restrict attention to prior distributions for contingency table models which are invariant under groups of permutations. Furthermore, we will restrict attention to invariant means and covariance structures, and therefore to distributions which are defined by these two quantities. Rather than considering an explicit parameterisation for an invariant model, we will focus on the prior mean and covariance for $\theta$ under the saturated model. Any non-saturated linear model for $\theta$ is then just a special case, with constraints on the prior mean and covariance. Suppose that the prior mean for $\theta$ is $\alpha$ and that the prior variance matrix is $\Sigma$. For a particular invariant model in realisation space, with projection matrix $Q$ onto the corresponding invariant subspace, $(I - Q)\alpha = 0$ and $(I - Q)\Sigma(I - Q) = 0$.

A natural prior distribution for $\theta$ is multivariate normal with mean $\mu$ and variance $\Sigma$.
where $1^T \mu = 0$ and $\Sigma 1 = 0$. The resulting prior distribution for $\mathbf{p}$ is logistic normal (see Aitchison, 1986, for details). The corresponding prior distribution for the vector of means $\mu$ in a Poisson model is the lognormal distribution. These prior distributions are more appropriate for the analysis of structured data using log-linear models, than the corresponding conjugate Dirichlet (or gamma) distributions, which have an insufficiently rich covariance structure.

For a square $r \times r$ table where $G = C_2 \times S_r$, one of the irreducible representations has multiplicity two, but all others have multiplicity one. Suppose that the irreducible representations $[r] \otimes [2]$, $[r-11] \otimes [2]$, $[r-11] \otimes [11]$, $[r-22] \otimes [2]$ and $[r-211] \otimes [11]$ are labelled 1, 2, 3, 4, 5 respectively. Then, from (??), a $G$-invariant variance matrix for a square contingency table must take the form

$$\Sigma = \frac{m_{111}}{r-1} v v^T + m_{211} Q_{21} + m_{222} Q_{22} + m_{311} Q_3 + m_{411} Q_4 + m_{511} Q_5 + m_{212} (T_{21} T_{22}^T + T_{22} T_{21}^T) \quad (6)$$

where $v, Q_{22}, Q_3, Q_4$ and $Q_5$ are defined in Section 5, $Q_{21} = I_r - 1/r J_r$ acting on the diagonal cells. and $T_{21}$ and $T_{22}$ are $r^2 \times (r-1)$ matrices whose columns are orthonormal bases for the two equivalent irreducible subspaces $V_{21}$ and $V_{22}$ described in Section 5.

The matrix $T_{21} T_{22}^T + T_{22} T_{21}^T$ is derived in Appendix A. It has non-zero entries only where the row cell is diagonal, and the column cell is off-diagonal, or vice versa. It takes the form

$$T_{21} T_{22}^T + T_{22} T_{21}^T = \frac{1}{\sqrt{2(r-2)}} \left( C - \frac{2}{r} D \right) \quad (7)$$

where the $r^2 \times r^2$ relationship matrices $C$ and $D$ are defined as as follows, denoting cells by row and column subscripts.

$$C_{ijkl} = \begin{cases} 
1 & \text{if } i = j \text{ and } k \neq l \text{ and } (i = k \text{ or } j = l) \\
1 & \text{if } i \neq j \text{ and } k = l \text{ and } (i = k \text{ or } j = l) \\
0 & \text{otherwise}
\end{cases}$$

$$D_{ijkl} = \begin{cases} 
1 & \text{if } i = j \text{ and } k \neq l \\
1 & \text{if } i \neq j \text{ and } k = l \\
0 & \text{otherwise}
\end{cases}$$

Each of these matrices represents relationships which are preserved under the action of $G$.

Projection matrices onto $G$-invariant subspaces are idempotent matrices of the form of (??). As discussed in Section 6, $m_{i11} = 0$ or 1 for $i = 1, 3, 4, 5$ and $M_2$ is a $2 \times 2$ projection matrix. The projection onto the irreducible invariant subspace $V_2(\phi)$, introduced in Section 5 corresponds to $m_{i11} = 0, i = 1, 3, 4, 5$ with $M_2$ given by (??). If $m_{311} = 1$ and $\phi = \arccos \sqrt{2/r}$, we have the projection onto $V_3 \oplus V_2(\arccos \sqrt{2/r})$, the parameter space for
the independence model for a $r \times r$ table, and $\Sigma = \mathbf{I}_r - \frac{1}{r^2} \mathbf{J}_r - (\mathbf{I}_r - \frac{1}{r} \mathbf{J}_r) \otimes (\mathbf{I}_r - \frac{1}{r} \mathbf{J}_r)$, as expected, when cells are ordered in the usual way.

## 4 Model Uncertainty

We have defined a $G$-invariant log-linear model as a $G$-invariant subspace of $\mathbf{1}_m^\perp$. As in Section 6, suppose that $\mathbf{1}_m^\perp = \bigoplus_{ij} V_{ij}$ is an irreducible decomposition of $V$ into $G$-invariant subspaces, where $i$ indexes inequivalent irreducible representations and $j$ runs from 1 to $e_i$, the multiplicity of the $i$th inequivalent irreducible component in $P_G$. This requires a particular choice of $V_{ij}$ when $e_i > 1$. Then, any non-saturated $G$-invariant model may be represented by a projection matrix of the form of (??), which in turn is defined by $\{M_i, i = 1, \ldots, l\}$, where each $M_i$ is an $e_i \times e_i$ projection matrix. Alternatively, we may represent a model by $\gamma \in \prod_{i=1}^l \{0, \ldots, e_i\}$ where $\gamma_i = \text{trace } M_i$ is the number of $i$-equivalent irreducible $G$-invariant subspaces in the model, together with a parameter $\phi_i$ for each $\gamma_i$ where $0 < \gamma_i < e_i$ specifying the $\gamma_i$ equivalent subspaces with respect to $V_{i1}, \ldots, V_{ie_i}$. Here $\gamma_i$ is the dimension of the subspace of $\mathbb{R}^{e_i}$ onto which $M_i$ projects, and $\phi_i$ describes that subspace. For example, if $e_i = 2$ and $\gamma_i = 1$, we have already described how, given a particular choice of $V_{i1}$ and $V_{i2}$, a scalar parameter $\phi$ can be used to define any of the equivalent $G$-invariant subspaces. The corresponding projection matrix $M_i$ is given by (??).

A log-linear model is therefore defined by $(\gamma, \phi_\gamma)$, where $\phi_\gamma$ is the collection of $\phi_i$ for those $i$ for which $0 < \gamma_i < e_i$. As statistical models are usually thought of as discrete entities, we will use $\gamma$ to define a ‘model’, and think of $\phi_\gamma$ as an extra parameter of the model. A model $\gamma$ is therefore a union of log-linear models, which is not itself log-linear, if $0 < \gamma_i < e_i$ for some $i$, as the set of resulting allowed values of $\theta$ is not a vector space. The exception to this is when the equivalent subspaces are one-dimensional. In this case, for each possible value of $\gamma_i > 1$, the union of all equivalent subspaces is equal to $\bigoplus_{j=1}^{\gamma_i} V_{ij}$. Therefore, when such a situation arises, the only values of $\gamma_i$ which need to be considered are $\gamma_i = 0$ and $\gamma_i = e_i$.

Typically, there is prior uncertainty about which log-linear model is appropriate for the data to be analysed. In a Bayesian analysis this is expressed through the prior distribution, together with uncertainty about model parameters. Here, we parameterise a non-saturated log-linear model $(\gamma, \phi_\gamma)$ using $\beta = \{\beta_{ij} = T_{ij}^T \theta, i = 1, \ldots, l; j = 1, \ldots, \gamma_i\}$, where the columns of the matrices $T_{ij}$ are an orthonormal basis for $(\gamma, \phi_\gamma)$, suitably constructed, as described in Section 6. Clearly $\beta$ depends on both $\gamma$ and $\phi_\gamma$, through the choice of the
matrices $T_{ij}$. The prior distribution for $(\beta, \phi\gamma, \gamma)$ is constructed hierarchically, as

$$f(\beta, \phi\gamma, \gamma) = f(\beta|\phi\gamma, \gamma)f(\phi\gamma|\gamma)f(\gamma).$$

(8)

A $G$-invariant prior distribution is ensured provided that $f(\beta|\phi\gamma, \gamma)$ is $G$-invariant for all $(\gamma, \phi\gamma)$. For distributions defined by prior means and covariances, this can be achieved by using the results discussed in Section 6. In particular, a $G$-invariant prior covariance for $\beta$ is of the form of (??). As all models under consideration are $g$-invariant, there is no restriction on the prior distribution for $(\gamma, \phi\gamma)$.

The posterior distribution for any function of $(\beta, \phi\gamma, \gamma)$ which has an interpretation across all models, $\theta$ for example, is ‘averaged’ with respect to the posterior distribution across models.

5 Example

Consider Table 3, analysed by Cazes (1990), which presents the distribution of marriages by lineage of each spouse amongst the Dogon of Boni, Mali. As exogamy (marriage outside the lineage) is prevalent, few of the observations lie on the diagonal. We will investigate the structure of this table using the $G$-invariant square table models introduced in Section 5.

<table>
<thead>
<tr>
<th>Husband’s lineage</th>
<th>Iariwa Ger.</th>
<th>Segiwa</th>
<th>Suraba</th>
<th>Pussuwoi</th>
<th>Iariwa</th>
<th>Tengo$_2$</th>
<th>Tengo</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iariwa Ger.</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Segiwa</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Suraba</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>Pussuwoi</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>33</td>
</tr>
<tr>
<td>Iariwa</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>16</td>
<td>42</td>
</tr>
<tr>
<td>Tengo$_2$</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td>14</td>
<td>51</td>
</tr>
<tr>
<td>Tengo</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>8</td>
<td>14</td>
<td>12</td>
<td>16</td>
<td>78</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>17</td>
<td>6</td>
<td>21</td>
<td>17</td>
<td>80</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>22</td>
<td>31</td>
<td>36</td>
<td>40</td>
<td>38</td>
<td>85</td>
<td>82</td>
<td>349</td>
</tr>
</tbody>
</table>

Table 1: Distribution of marriages in the village of Nemgéné by lineage of each spouse

For a square table, with four representations, $i = 1, 3, 4, 5$, of multiplicity 1 and one, $i = 2$ of multiplicity 2, there are 48 possible models $\gamma$, and a single parameter $\phi$ is present when $\gamma_2 = 1$. In the absence of any strong prior information about which models are most
appropriate for these data, we give all models with $\gamma_2 = 1$ prior probability $1/32$, and all other models prior probability $1/64$. We give $\phi$ a uniform prior distribution over $(-\pi/2, \pi/2]$. Models with $\gamma_2 = 1$ are given higher probability because, for example, if $V_{21}$ and $V_{22}$ were inequivalent, and all invariant models were given equal prior probability, the prior probability of $\{V_{21}, V_{22}\}$ would be twice that of $\emptyset$ or $V_{21} \oplus V_{22}$.

Here, a proper, but diffuse prior for $\theta$ in the saturated model is specified by setting $\alpha = 0$, $m_{i11} = \psi'(\lambda), i = 1, \ldots, 5$ and $m_{212} = m_{221} = 0$ in (??), and again we choose $\lambda = 1/2$. Again the appropriate prior variance matrix for $\theta$ in any non-saturated model is the marginal prior variance in the appropriate subspace. Here, this involves setting $m_{i11} = 0$ when $\gamma_i = 0$, and if $\gamma_2 = 1$, then

$$M_2 = \psi'(\lambda) \begin{pmatrix} \cos^2 \phi & \cos \phi \sin \phi \\ \cos \phi \sin \phi & \sin^2 \phi \end{pmatrix}.$$  

Suppose that the columns of $T_1, T_{21}, T_{22}, T_3, T_4$ and $T_5$ are specified orthonormal bases of the corresponding invariant subspaces. Then, as in Section 6, any invariant model can be parameterised using $\beta = T^T \theta$, where $T$ has the columns of the appropriate submatrices (if $\gamma_2 = 1$, then $T$ contains the columns of $\cos(\phi)T_{21} + \sin(\phi)T_{22}$). Then $\theta = T \beta$, where $T$ depends on $\gamma$, and on $\phi$ if $\gamma_2 = 1$. However, for any invariant model parameterised this way $\text{Var}(\beta) = \psi'(\lambda)I$.

Posterior marginal densities may be computed using Gibbs sampling. When $\gamma_2 \neq 1$, then for any linear parameterisation of an invariant subspace, such as $\beta$, the joint posterior density is log-concave. One consequence of this is that univariate conditional posterior densities are log-concave, and therefore adaptive rejection sampling (Gilks and Wild, 1992) may be used to generate observations from these univariate distributions. Posterior correlations are not excessive, and the procedure seems to be efficient and reliable. When $\gamma_2 = 1$, the extra parameter $\phi$ needs to be considered. Conditional on a particular value of $\phi$, the model is still log-linear, so the posterior distribution of $\beta$ is still log-concave. A Metropolis step can be used to generate $\phi$ conditional on $\beta$. In this paper, we focus on calculating posterior probabilities for invariant models. These probabilities are the mixing weights for model-averaged posterior inference. In order to calculate the marginal likelihood for a particular model $\gamma$, we use Laplace’s method. When $\gamma_2 = 1$, we use Laplace’s method to integrate out $\beta$ at a series of fixed values of $\phi$, and then use a simple trapezium rule to marginalise over $\phi$.

We used these methods to analyse Table 3. The log marginal likelihood for each model with $\gamma_2 = 1$, as a function of $\phi$, is displayed in Figure 1. Exponentiating each of these curves gives an unnormalised posterior density for $\phi$, from which the overall marginal likelihood for the model, integrated over $\phi$, can be obtained. The shape of the marginal posterior density for $\phi$ depends on whether $V_1$ is also included in the model. The posterior mode for $\phi$ is at
around $0.25\pi$ for models which include $V_1$ and at around $0.36\pi$ for models without this term. Note that for this table, $\arccos \sqrt{2/r} = \pi/3$. The most probable model displayed in Figure 1 is more than 12500 times more probable than any of the other models, when $\phi$ is integrated out. Indeed, this model, $\gamma = (1,1,0,0,0)$, is the most probable of all models considered, and has a posterior probability of 0.9867. The only other model with non-negligible probability is $\gamma = (1,2,0,0,0)$, which includes both $V_{21}$ and $V_{22}$, and has a posterior probability of 0.0132. This model reflects quasi-independence of off-diagonal cells, together with marginal homogeneity. The diagonal and off-diagonal cell probabilities are modelled independently. In standard log-linear model notation, this may be expressed as

$$\log p_{ij} = \begin{cases} \mu + \alpha_i + \alpha_j & i \neq j \\ \mu + \delta & i = j. \end{cases}$$

The preferred model, $\gamma = (1,1,0,0,0)$, reflects some common structure of the diagonal and off-diagonal cells. It may be expressed as

$$\log p_{ij} = \begin{cases} \mu + (\alpha_i + \alpha_j) \sin \phi / \sqrt{2(r - 2)} & i \neq j \\ \mu + \delta + \alpha_i \cos \phi & i = j. \end{cases}$$

Hence $\phi$ represents the uniformity of the diagonal cells relative to the off-diagonal margins. If $\phi = 0$ then the model represents off-diagonal marginal uniformity, whereas if $\phi = \pi/2$, the model represents diagonal uniformity. Intermediate values of $\phi$ lie between these extremes, with $\phi = \arccos \sqrt{2/r}$, the diagonal parameter model, representing equivalent diagonal and off-diagonal behaviour. For Table 3, the modal value of $\phi$ is between 0 and $\arccos \sqrt{2/r}$, indicating stronger off-diagonal marginal uniformity.

Model averaged inferences will primarily be based on $\gamma = (1,1,0,0,0)$, (averaged over $\phi$) with a small contribution from $\gamma = (1,2,0,0,0)$. We do not present any such inferences here, but they may easily be obtained using Markov chain Monte Carlo, as described above.

6 Discussion

The two examples presented illustrate how the consideration of invariance under an appropriate permutation group can be used to construct a reference set of models for analysis. Often the models are familiar, but in other cases, such as the most probable model in Section 10.2, less so. Invariant prior distributions for model parameters, required to properly reflect prior ignorance follow easily.

The model-averaged prior predictive distribution is permutation invariant, as we have only considered permutation invariant models, and permutation invariant prior distributions.
Figure 1: Log marginal likelihoods for each model with $\gamma_2 = 1$, plotted as a function of $\phi/\pi$. Models which include $V_1$ are represented by solid lines, those without by dashed lines. For each of these two groups the models illustrated include, in order of modal height: $\emptyset, V_3, V_4, V_3 \oplus V_4, V_3 \oplus V_5, V_4 \oplus V_5, V_3 \oplus V_4 \oplus V_5$

for model parameters. It is possible to achieve a permutation invariant prior predictive distribution with a much less restrictive choice of models. For example, the outlier models described by Albert (1997), which allow specified cells to deviate from a model such as row-column independence in a two-way table, are clearly not permutation invariant, but a symmetric mixture of such models results in a permutation invariant prior predictive distribution.

The approach envisaged in this paper can be considered to be largely nonparametric, with estimation or prediction averaged across models, being the main concern. In particular, we have not been greatly concerned with how individual models might be interpreted. It is clear that permutation invariance alone, does not necessarily result in models which have a straightforward interpretation. In particular, all log-linear interaction models for a multiway
table are invariant under the product symmetric group. Usually, attention is restricted to the hierarchical models, which can be easily interpreted in terms of (conditional) association. In situations where the appropriate group is symmetric, or a product of symmetric groups, McCullagh (1998) considers a form of invariance which is more restrictive than permutation invariance. He defines a (log)-linear model as a rule which assigns subspaces across situations where the sets of levels $I_c$ of the factors $c$ change, and in particular where $|I_c|$ is not constant. He considers invariance under various morphisms between tables, which may be interpreted as selecting, or merging elements of $I_c$. For a multiway table, only the hierarchical log-linear interaction models are then invariant. In other examples, the sets of models are also reduced.

The forms of invariance considered by McCullagh are most easily interpreted when the factors are explanatory variables, rather than for multinomial models, where the factors comprise the response. However, the concept of selection of factor levels may still be interpreted in terms of restriction of the response to certain levels, and analysis conditional on these levels. The appropriate $\theta$ for the conditional analysis may be obtained by a simple linear transformation of the original $\theta$. This is similar to the way the diagonal and off-diagonal cell probabilities were considered in the analysis of a square table. Hence, in certain circumstances the set of invariant models may be reduced further. This is the subject of ongoing research, particularly with respect to which prior distributions are then appropriate.

References


Appendix A

We require to find $T_{21}T_{22}^T + T_{22}T_{21}^T$, where the columns of $T_{21}$ and $T_{22}$ are orthonormal bases for $V_{21}$ and $V_{22}$ respectively.

Note that $V_{21}$ and $V_{22}$ are isomorphic to $1_r^\perp$. Suppose that $T$ is any $r \times (r-1)$ matrix whose columns are an orthonormal basis for $1_r^\perp$. Therefore $TT^T = I_r - 1/r J_r$, $T^TT = I_{r-1}$, $T^T1_r = 0_{r-1}$ and $T1_{r-1} = 0_r$.

$T_{21}$ and $T_{22}$ must be equivalent under permutation. This can be ensured by setting $T_{21} = b_1 B_1 T$ and $T_{22} = b_2 B_2 T$, where $B_1$ and $B_2$ are $r^2 \times r$ matrices whose columns are equivalent spanning sets for $V_{21}$ and $V_{22}$ respectively. An obvious choice for $B_1$ is the submatrix of $Q_{21}$ obtained by extracting the non-zero columns, which correspond to diagonal cells, and placing them in the order $1, \ldots, r$. An equivalent $B_2$ can then be obtained from $Q_{22}$. The $i$th column of $B_2$ is the sum of all columns of $Q_{22}$ which correspond to cells in the $i$th row or $i$th column of the table. Hence,

$$B_{2ijk} = \begin{cases} 1 - \frac{2}{r} & \text{if } i = k \text{ or } j = k \\ -\frac{2}{r} & \text{otherwise.} \end{cases}$$

Now, for the columns of $T_{21}$ and $T_{22}$ to be orthonormal, we require $b_i^2 T_i^T B_i^T B_i T = I_{r-1}$ for $i = 1, 2$. Hence $b_1 = 1$ and $b_2 = 1/\sqrt{2(r-2)}$. Then,

$$T_{21}T_{22}^T + T_{22}T_{21}^T = \frac{1}{\sqrt{2(r-2)}} (B_1TT^T B_2^T + B_2TT^T B_1^T),$$

which leads to (??).