UNIVERSITY OF
Southampton
Southampton Statistical
Sciences Research Institute

Working Paper M09/11
Methodology

Bayesian Inference for Poisson and Multinomial Log-linear Models

Jonathan J. Forster

Abstract

Categorical data frequently arise in applications in the social sciences. In such applications,the class of log-linear models, based on either a Poisson or (product) multinomial response distribution, is a flexible model class for inference and prediction. In this paper we consider the Bayesian analysis of both Poisson and multinomial log-linear models. It is often convenient to model multinomial or product multinomial data as observations of independent Poisson variables. For multinomial data, Lindley (1964) showed that this approach leads to valid Bayesian posterior inferences when the prior density for the Poisson cell means factorises in a particular way. We develop this result to provide a general framework for the analysis of multinomial or product multinomial data using a Poisson log-linear model. Valid finite population inferences are also available, which can be particularly important in modelling social data.We then focus particular attention on multivariate normal prior distributions for the log-linear model parameters.Here, an improper prior distribution for certain Poisson model parameters is required for valid multinomial analysis, and we derive conditions under which the resulting posterior distribution is proper.We also consider the construction of prior distributions across models, and for model parameters, when uncertainty exists about the appropriate form of the model. We present classes of Poisson and multinomial models, invariant under certain natural groups of permutations of the cells. We demonstrate that, if prior belief concerning the model parameters is also invariant, as is the case in a `reference' analysis, then choice of prior distribution is considerably restricted. The analysis of multivariate categorical data in the form of a contingency table is considered in detail. We illustrate the methods with two examples.

Bayesian Inference for Poisson and Multinomial Log-linear Models

Jonathan J. Forster[1]

SUMMARY

Categorical data frequently arise in applications in the Social Sciences. In such applications,the class of log-linear models, based on either a Poisson or (product) multinomial response distribution, is a flexible model class for inference and prediction. In this paper we consider the Bayesian analysis of both Poisson and multinomial log-linear models. It is often convenient to model multinomial or product multinomial data as observations of independent Poisson variables. For multinomial data, Lindley (1964) showed that this approach leads to valid Bayesian posterior inferences when the prior density for the Poisson cell means factorises in a particular way. We develop this result to provide a general framework for the analysis of multinomial or product multinomial data using a Poisson log-linear model. Valid finite population inferences are also available, which can be particularly important in modelling social data. We then focus particular attention on multivariate normal prior distributions for the log-linear model parameters. Here, an improper prior distribution for certain Poisson model parameters is required for valid multinomial analysis, and we derive conditions under which the resulting posterior distribution is proper. We also consider the construction of prior distributions across models, and for model parameters, when uncertainty exists about the appropriate form of the model. We present classes of Poisson and multinomial models, invariant under certain natural groups of permutations of the cells. We demonstrate that, if prior belief concerning the model parameters is also invariant, as is the case in a 'reference' analysis, then choice of prior distribution is considerably restricted. The analysis of multivariate categorical data in the form of a contingency table is considered in detail. We illustrate the methods with two examples.

1 Introduction

Suppose that in each of c groups, N_i $(i = 1, \ldots, c)$ individuals are independently classified into one of n_i $(i = 1, \ldots, c)$ categories. Therefore there are a total of $\sum_{i=1}^{c} N_i$ individuals

[1] School of Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK
email: J.J.Forster@soton.ac.uk

classified into a total of $n = \sum_{i=1}^{c} n_i$ categories. The observed data can be represented as a vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ of n cell counts, which is subject to the constraint $\boldsymbol{Cy} = \boldsymbol{N} = (N_1, \ldots, N_c)^T$, where $\boldsymbol{C}$ is a $c \times n$ matrix, with the property that every column contains $c-1$ zeros, the remaining element being equal to one. Hence if $c = 1$, $\boldsymbol{C}$ is a row vector of ones, and the constraint is the usual simple multinomial constraint of a fixed grand total.

The cells are therefore divided into c non-overlapping strata, with each stratum total fixed in advance. The likelihood for this product multinomial model is

$$f(\boldsymbol{y}|\boldsymbol{p}) \quad \propto \quad \prod_{i=1}^{n} p_i^{y_i}, \tag{1}$$

where $\boldsymbol{Cp} = \boldsymbol{1}_c$. Therefore, unless further constraints are placed on $\boldsymbol{p}$, there are $n - c$ free parameters. We will refer to this model as a multinomial model for any $c > 0$.

Even when they are fixed in advance, by design, it is often convenient to treat $N_1, \ldots, N_c$ as observations of independent Poisson random variables, in which case $y_1, \ldots, y_n$ are also Poisson, and independent. The distribution of $\boldsymbol{y}$ is then represented by the corresponding vector of cell means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$.

The likelihood for the Poisson model is

$$f(\boldsymbol{y}|\boldsymbol{\mu}) \quad \propto \quad \exp\left(-\sum_{i=1}^{n} \mu_i\right) \prod_{i=1}^{n} \mu_i^{y_i}. \tag{2}$$

An alternative parameterisation for the Poisson model is through $\boldsymbol{\mu}^+ = \boldsymbol{C\mu}$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$, where $\pi_i = \mu_i/[\boldsymbol{C}^T\boldsymbol{\mu}^+]_i$. Now, the Poisson likelihood can be written as

$$f(\boldsymbol{y}, \boldsymbol{N}|\boldsymbol{\mu}^+, \boldsymbol{\pi}) \quad \propto \quad \exp\left(-\sum_{i=1}^{c} \mu_i^+\right) \prod_{i=1}^{c} \mu_i^{+N_i} \prod_{i=1}^{n} \pi_i^{y_i} \tag{3}$$

where $\boldsymbol{C\pi} = \boldsymbol{1}_c$ and $\boldsymbol{Cy} = \boldsymbol{N}$. This is simply a result of the familiar factorisation $f(\boldsymbol{y}, \boldsymbol{N}|\boldsymbol{\mu}^+, \boldsymbol{\pi}) = f(\boldsymbol{N}|\boldsymbol{\mu}^+)f(\boldsymbol{y}|\boldsymbol{N}, \boldsymbol{\pi})$. As described above, this is equivalent to the stratum totals $\boldsymbol{N}$ being drawn from independent Poissons, mean $\boldsymbol{\mu}^+$, and then conditional on $\boldsymbol{N}$, the cell counts $\boldsymbol{y}$ have a product multinomial distribution.

The Poisson model is easier to deal with, particularly using standard software, as the parameter $\boldsymbol{\mu}$ is unconstrained (apart from positivity), and the cell counts are observations of independent variables. Baker (1994) provides a series of examples. Lang (1996) considers likelihood-based inference for the parameters of log-linear models, and describes how inferences for multinomial or product multinomial models may be obtained from inferences for Poisson models, extending results of Birch (1963). Here, we describe how this applies to Bayesian inference. Similar advantages accrue. For example, Markov chain Monte Carlo methods for posterior inference are typically much more straightforward to apply to Poisson models than to multinomial models.

2 Bayesian Inference

When data are observed using a product multinomial sampling scheme, the appropriate Bayesian inference is obtained by specifying a prior distribution for $\boldsymbol{p}$, and obtaining the posterior density for $\boldsymbol{p}$. Hence, $f(\boldsymbol{p}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{p})f(\boldsymbol{p})$, and therefore, from (1),

$$f(\boldsymbol{p}|\boldsymbol{y}) \quad \propto \quad f(\boldsymbol{p}) \prod_{i=1}^{n} p_i^{y_i}, \tag{4}$$

where $\boldsymbol{C}\boldsymbol{p} = \mathbf{1}_c$.

If, instead, the Poisson likelihood is assumed, then a prior must be specified for $\boldsymbol{\mu}$, or equivalently $(\boldsymbol{\mu}^+, \boldsymbol{\pi})$, and the resulting posterior density is

$$f(\boldsymbol{\mu}^+, \boldsymbol{\pi}|\boldsymbol{y}) \quad \propto \quad f(\boldsymbol{N}|\boldsymbol{\mu}^+) f(\boldsymbol{y}|\boldsymbol{N}, \boldsymbol{\pi}) f(\boldsymbol{\mu}^+, \boldsymbol{\pi}),$$

as the likelihood (3) factorises. Therefore it is clear that if $\boldsymbol{\mu}^+$ and $\boldsymbol{\pi}$ are *a priori* independent, and hence $f(\boldsymbol{\mu}^+, \boldsymbol{\pi}) = f(\boldsymbol{\mu}^+)f(\boldsymbol{\pi})$, then they will also be *a posteriori* independent. In particular, the marginal posterior density for $\boldsymbol{\pi}$ will be given by

$$f(\boldsymbol{\pi}|\boldsymbol{y}) \quad \propto \quad f(\boldsymbol{\pi}) \prod_{i=1}^{n} \pi_i^{y_i}, \tag{5}$$

The equivalence between (4) and (5) allows us to use the more convenient Poisson representation to analyse product multinomial data. The required posterior distribution for $\boldsymbol{p}$ can be obtained by transforming $\boldsymbol{\mu}$ to $\boldsymbol{\pi}$. When a Monte Carlo approach is being used, this is especially straightforward. All that is required is to specify the prior for $\boldsymbol{\mu}$ correctly. An appropriate prior for $\boldsymbol{\mu}$ leads to a prior for $(\boldsymbol{\mu}^+, \boldsymbol{\pi})$ with the properties that (i) $\boldsymbol{\mu}^+$ and $\boldsymbol{\pi}$ are independent; and (ii) the prior for $\boldsymbol{\pi}$ is the required prior distribution for the product multinomial parameter $\boldsymbol{p}$.

Apart from the independence constraint, any choice of prior for $\boldsymbol{\mu}^+$ will suffice. Lindley (1964) made use of this result in the multinomial case ($c = 1$). The most straightforward example is where $\mu_1, \ldots, \mu_n$ have independent gamma distributions with corresponding shape parameters $\alpha_1, \ldots, \alpha_n$, and common scale parameter β. Then

$$f(\boldsymbol{\mu}) \quad \propto \quad \exp\left(-\beta \sum_{i=1}^{n} \mu_i\right) \prod_{i=1}^{n} \mu_i^{\alpha_i - 1}.$$

The Jacobian for the transformation from $\boldsymbol{\mu}$ to $(\boldsymbol{\mu}^+, \boldsymbol{\pi})$ is $\prod_{i=1}^{c} \mu_i^{+c_i - 1}$. Therefore,

$$f(\boldsymbol{\mu}^+, \boldsymbol{\pi}) \quad \propto \quad \exp\left(-\beta \sum_{i=1}^{c} \mu_i^+\right) \prod_{i=1}^{c} \mu_i^{+[\boldsymbol{C}\boldsymbol{\alpha}]_i - 1} \prod_{i=1}^{n} \pi_i^{\alpha_i - 1}.$$

Hence $\boldsymbol{\mu}^+$ and $\boldsymbol{\pi}$ are independent, and the marginal prior distribution for $\boldsymbol{\pi}$ is a product Dirichlet distribution. Hence, posterior inference for (product) multinomial data, with a (product) Dirichlet prior may be obtained by using an independent Poisson likelihood, and appropriate gamma priors.

The result is also applicable when $\boldsymbol{y}$ is drawn from a finite population with corresponding population frequencies $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Here we assume that the stratum population totals $\boldsymbol{CY} = \boldsymbol{Y}^+ = (Y_1^+, \ldots, Y_c^+)^T$ are known. Where $c = 1$ (multinomial sampling), Ericson (1969) proposed a prior distribution for $\boldsymbol{Y}$ reflecting exchangeability of the units comprising the population. This can be adapted easily to exchangeability of the units within each stratum. Following Ericson (1969), the prior is constructed as a two stage hierarchical distribution with a (product) multinomial$(\boldsymbol{Y}^+, \boldsymbol{p})$ distribution for $\boldsymbol{Y}|\boldsymbol{p}$ at the first stage, where $\boldsymbol{Cp} = \boldsymbol{1}_c$. The hyperparameter $\boldsymbol{p}$ is then given an arbitrary second stage distribution $f(\boldsymbol{p})$. The resulting posterior for the unsampled population cell frequencies is

$$\boldsymbol{Y} - \boldsymbol{y}|\boldsymbol{p} \sim \text{multinomial}(Y^+ - N, \boldsymbol{p})$$

with $f(\boldsymbol{p})$ updated to $f(\boldsymbol{p}|\boldsymbol{y})$ in the posterior, using (4). As we can obtain $f(\boldsymbol{p}|\boldsymbol{y})$ by assuming a Poisson sampling scheme with an appropriate prior, it is clear that the the corresponding finite population inferences will also be available.

In some situations there may exist a number of plausible models for $\boldsymbol{p}$, which constrain $\boldsymbol{p}$ so that its effective dimension is less than $n - c$. In this situation, the prior is constructed to reflect this. Suppose that the possible models are denoted by $m \in \{1, \ldots, M\}$, and the joint prior distribution of $(m, \boldsymbol{p})$ is of the form $f(m)f(\boldsymbol{p}|m)$ where $f(\boldsymbol{p}|m)$ places all its mass on values of $\boldsymbol{p}$ constrained in a way consistent with m. Then,

$$f(m, \boldsymbol{p}|\boldsymbol{y}) \;\propto\; f(\boldsymbol{y}|\boldsymbol{p})f(m)f(\boldsymbol{p}|m), \qquad m \in \{1, \ldots, M\}.$$

The equivalent Poisson models constrain $\boldsymbol{\pi}$, and assuming that the prior for $\boldsymbol{\mu}$ satisfies conditions (i) and (ii) above *for all* m, we have

$$f(m, \boldsymbol{\pi}, \boldsymbol{\mu}^+|\boldsymbol{y}) \;\propto\; f(\boldsymbol{N}|\boldsymbol{\mu}^+)f(\boldsymbol{y}|\boldsymbol{N}, \boldsymbol{p})f(m)f(\boldsymbol{\pi}|m)f(\boldsymbol{\mu}^+|m), \qquad m \in \{1, \ldots, M\}.$$

Now, for the marginal posterior distribution of $(m, \boldsymbol{\pi})$ under the Poisson model to be identical to the posterior distribution of $(m, \boldsymbol{p})$ under the multinomial model, we require an extra condition, that $f(\boldsymbol{\mu}^+|m)$ does not depend on m. The same prior for $\boldsymbol{\mu}^+$ is required for all models. As the models relate to $\boldsymbol{p}$ and $\boldsymbol{\pi}$, this does not seem to be a serious restriction. Gûnel and Dickey (1974) consider the Bayes factor for comparing independence and saturated models in a two-way contingency table, and give an example where inference under Poisson and multinomial models differs when this condition is violated.

3 Log-linear models

Often, the categories $1, \ldots, n$, arise as a result of a cross-classification of individuals by a number of categorical variables. The resulting data form a contingency table, and it is common to investigate the structure of the table using log-linear models for $\boldsymbol{p}$ or $\boldsymbol{\mu}$. The saturated log-linear model for $\boldsymbol{\mu}$ allows $\log \boldsymbol{\mu}$ to take any value in $\mathcal{R}^n$. A non-saturated model constrains $\log \boldsymbol{\mu}$ to lie in some vector subspace of $\mathcal{R}^n$.

Let $\boldsymbol{\theta}$ be the multivariate stratum-centred logit

$$\boldsymbol{\theta} \;=\; \log \boldsymbol{p} \;-\; \boldsymbol{C}^T \mathrm{diag}(\boldsymbol{n})^{-1} \boldsymbol{C} \log \boldsymbol{p}$$

and hence

$$\log \boldsymbol{p} \;=\; \boldsymbol{\theta} \;-\; \boldsymbol{C}^T \log(\boldsymbol{C} \exp \boldsymbol{\theta}).$$

For Poisson models, the equivalent logit is defined as

$$\boldsymbol{\theta} \;=\; \log \boldsymbol{\pi} \;-\; \boldsymbol{C}^T \mathrm{diag}(\boldsymbol{n})^{-1} \boldsymbol{C} \log \boldsymbol{\pi} \;=\; \log \boldsymbol{\mu} \;-\; \boldsymbol{C}^{\mathrm{T}} \mathrm{diag}(\boldsymbol{n})^{-1} \boldsymbol{C} \log \boldsymbol{\mu}.$$

Therefore $\boldsymbol{C\theta} = \boldsymbol{0}_c$, as $\boldsymbol{C}\boldsymbol{C}^T = \mathrm{diag}(\boldsymbol{n})$, and $\boldsymbol{\theta}$ lies in $N(\boldsymbol{C})$, a $(n-c)$-dimensional vector subspace of $\mathcal{R}^n$. Indeed, $\boldsymbol{\theta}$ is the orthogonal projection of $\log \boldsymbol{p}$ or $\log \boldsymbol{\mu}$ onto $N(\boldsymbol{C})$. This is a much more convenient parameter space to deal with than $\{\boldsymbol{p} : p_i > 0, i = 1, \ldots, n; \boldsymbol{Cp} = \boldsymbol{1}_c\}$, the equivalent parameter space for $\boldsymbol{p}$. If $c = 1$ then $\boldsymbol{\theta} = \log \boldsymbol{p} - \log g(\boldsymbol{p})$ is the centred logratio used by Aitchison (1986, p79) where $g(\boldsymbol{p})$ is the geometric mean of $\{p_1, \ldots, p_n\}$. Any alternative multivariate logit may be obtained from $\boldsymbol{\theta}$ by linear transformation.

We define a log-linear model for $\boldsymbol{p}$ to be any vector subspace of $N(\boldsymbol{C})$. We express the model as $\boldsymbol{\theta} = \boldsymbol{X\beta}$, where $\boldsymbol{X}$ is a $n \times p$ matrix, and $\boldsymbol{CX} = \boldsymbol{0}$. Therefore the saturated product-multinomial model is $R(\boldsymbol{X})$ for any $n \times (n-c)$ $\boldsymbol{X}$ whose columns span $N(\boldsymbol{C})$. The Poisson log-linear model equivalent to $R(\boldsymbol{X})$ is $R(\boldsymbol{Z})$ where $\boldsymbol{Z} = (\boldsymbol{X} \;\, \boldsymbol{C}^T)$, which can be expressed as $\log \boldsymbol{\mu} = \boldsymbol{X\beta} + \boldsymbol{C}^T \boldsymbol{\phi}$. Here, $\boldsymbol{\phi} = \mathrm{diag}(\boldsymbol{n})^{-1} \boldsymbol{C} \log \boldsymbol{\mu}$. This constrains $\boldsymbol{\pi}$ in exactly the same way that the product multinomial model constrains $\boldsymbol{p}$, so the likelihoods for $\boldsymbol{\beta}$, $f(\boldsymbol{y}|\boldsymbol{N}, \boldsymbol{\beta})$, are identical under the two models.

The posterior density under the Poisson model is

$$f(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{y}) \;\;\propto\;\; f(\boldsymbol{N}|\boldsymbol{\mu}^+) f(\boldsymbol{y}|\boldsymbol{N}, \boldsymbol{\beta}) f(\boldsymbol{\beta}, \boldsymbol{\phi})$$

where $f(\boldsymbol{\beta}, \boldsymbol{\phi})$ is the joint prior density for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, and

$$\log \boldsymbol{\mu}^+ \;=\; \boldsymbol{\phi} + \log(\boldsymbol{C} \exp \boldsymbol{X\beta}).$$

Transforming $(\boldsymbol{\beta}, \boldsymbol{\phi})$ to $(\boldsymbol{\beta}, \log \boldsymbol{\mu}^+)$, we obtain

$$f(\boldsymbol{\beta}, \log \boldsymbol{\mu}^+|\boldsymbol{y}) \;\;\propto\;\; f(\boldsymbol{N}|\boldsymbol{\mu}^+) f(\boldsymbol{y}|\boldsymbol{N}, \boldsymbol{\beta}) f(\boldsymbol{\beta}, \boldsymbol{\phi}\{\boldsymbol{\mu}^+, \boldsymbol{\beta}\})$$

as the Jacobian for the transformation from $(\boldsymbol{\beta}, \boldsymbol{\phi})$ to $(\boldsymbol{\beta}, \log \boldsymbol{\mu}^+)$ is one. A sufficient condition for marginal posterior inference for $\boldsymbol{\beta}$ from this model to be equivalent to the multinomial model is that

$$f(\boldsymbol{\beta}, \boldsymbol{\phi}) \quad = \quad f(\boldsymbol{\beta}) \tag{6}$$

where $f(\boldsymbol{\beta})$ is the required prior density for $\boldsymbol{\beta}$. This is an improper prior which is uniform over $\mathcal{R}^c$ for $\boldsymbol{\phi}$. The posterior distribution will still be proper, unless one of the strata has no observations, in which case this stratum can be eliminated from the analysis as the corresponding cells are structural zeros. The proof of this condition appears in Section 6.

The most straightforward prior for $\boldsymbol{\beta}$ under the product multinomial model is a multivariate normal distribution for $\boldsymbol{\beta}$. This results in a lognormal distribution for $\log \boldsymbol{\mu}$ and a logistic normal distribution for $\boldsymbol{p}$ (normal for any multivariate logit; see Aitchison, 1986, for details). King and Brooks (2001) derive the relationship between the distributions of $\boldsymbol{\beta}$, $\log \boldsymbol{\mu}$ and $\boldsymbol{p}$ for a particular model matrix. Suppose that we assume a Poisson model, with a multivariate normal prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\phi})$ with mean $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_\beta, \boldsymbol{\alpha}_\phi)$ and precision (inverse variance) matrix $\boldsymbol{S}$, partitioned

$$\boldsymbol{S} \quad = \quad \begin{pmatrix} \boldsymbol{S}_{\beta\beta} & \boldsymbol{S}_{\beta\phi} \\ \boldsymbol{S}_{\phi\beta} & \boldsymbol{S}_{\phi\phi} \end{pmatrix}.$$

Knuiman and Speed (1988) showed that, if $\boldsymbol{S}_{\beta\phi} = \boldsymbol{S}_{\phi\beta}^T = \boldsymbol{0}$ and $\boldsymbol{S}_{\phi\phi} = \boldsymbol{0}$, which completely eliminates $\boldsymbol{\mu}^+$ (or $\boldsymbol{\phi}$) from the prior, then the posterior mode for $\boldsymbol{\beta}$ and posterior dispersion, calculated as negative second derivative of log posterior density at the posterior mode, are the same for Poisson and (product)-multinomial models. In fact, as this prior is of the form (6), the entire posterior density for $\boldsymbol{\beta}$ is identical under the two sampling models. The marginal prior for $\boldsymbol{\beta}$ is proper provided that $\boldsymbol{S}_{\beta\beta}$ is positive definite.

Usually, there exists uncertainty about which log-linear model is appropriate for the data. Suppose that the possible log-linear models are denoted by $m \in \{1, \ldots, M\}$, where model m is $\boldsymbol{\theta} = \boldsymbol{X}_m \boldsymbol{\beta}^m$ (multinomial) and $\log \boldsymbol{\mu} = \boldsymbol{X}_m \boldsymbol{\beta}^m + \boldsymbol{C}^T \boldsymbol{\phi}$ (Poisson). If, for the Poisson model, $f(m, \boldsymbol{\beta}^m, \log \boldsymbol{\mu}^+) = f(m) f(\boldsymbol{\beta}^m | m)$ then the prior distribution for $\boldsymbol{\phi}$ (or $\log \boldsymbol{\mu}^+$) is uniform for each model. The resulting posterior density is

$$f(m, \boldsymbol{\beta}^m, \log \boldsymbol{\mu}^+ | \boldsymbol{y}) \quad \propto \quad f(\boldsymbol{N} | \boldsymbol{\mu}^+) f(\boldsymbol{y} | \boldsymbol{\beta}^m) f(m) f(\boldsymbol{\beta}^m | m).$$

As $\boldsymbol{\mu}^+$ and $(m, \boldsymbol{\beta}^m)$ are *a posteriori* independent, the marginal distribution for $(m, \boldsymbol{\beta}^m)$, is the same as for the corresponding multinomial analysis, where $\boldsymbol{\mu}^+$ is absent. In particular relative marginal likelihoods of models (Bayes factors) are the same under the multinomial model. Although the prior distribution for $\boldsymbol{\phi}$ is improper, the same improper prior appears in all models.

4 Permutation Invariant Models

Under model uncertainty, we require a prior distribution $f(m)$ over the set M of all possible log-linear models. As we define log-linear models as vector subspaces of $\mathcal{R}^n$ for Poisson sampling or $N(\boldsymbol{C})$ for multinomial sampling, the set M is potentially infinite. This set may be reduced by considering only those models which are invariant under certain permutations of the category labels $\{1, \ldots, n\}$. This is desirable in any situation where the prior belief about the cell probabilities is unaltered when the cell labels are permuted in certain ways. Even in cases where the permutations are 'too restrictive' and overstate the degree of prior uncertainty, the set of models obtained may still be suitable for a reference analysis.

We denote a permutation under which invariance is required by g, and the corresponding $n \times n$ permutation matrix, acting on $\boldsymbol{y}$, $\boldsymbol{\mu}$ or $\boldsymbol{\theta}$ by $\boldsymbol{P}_g$. The set of all such permutations forms a group G under composition. As log-linear models are vector subspaces of $\mathcal{R}^n$, determining models which are invariant under G is equivalent to finding G-invariant subspaces of $\mathcal{R}^n$, in other words to finding subspaces $V_i \subset \mathcal{R}^n$ such that $\boldsymbol{P}_g \log \boldsymbol{\mu} \in V_i$ for all $\log \boldsymbol{\mu} \in V_i$ and all $g \in G$.

When $c > 0$, and certain stratum totals are fixed in advance, it clearly does not make sense to consider permutations which alter the strata. Hence, any two cells are in the same stratum after permutation if and only if they were originally in the same stratum. This is equivalent to requiring that $N(\boldsymbol{C})$ is itself an invariant subspace of $\mathcal{R}^n$ under any permutation being considered. The stratum-centred logit $\boldsymbol{\theta}$ is then invariant under any strata-preserving permutation g, in the sense that $\boldsymbol{\theta}(\boldsymbol{P}_g \boldsymbol{p}) = \boldsymbol{P}_g \boldsymbol{\theta}(\boldsymbol{p})$. For example, for simple multinomial sampling, where $\boldsymbol{C} = (1, \cdots, 1)$, then clearly $N(\boldsymbol{C})$ is invariant under any permutation. In the following, we shall therefore restrict attention to Poisson log-linear models, and consider invariant subspaces of $\mathcal{R}^n$. For the same set of permutations, invariant multinomial log-linear models are simply those invariant subspaces of $\mathcal{R}^n$, which are also invariant subspaces of $N(\boldsymbol{C})$.

Determination of the G-invariant subspaces of $\mathcal{R}^n$ utilises group representation theory. See, for example James and Liebeck (1993) or, for applications in Statistics, Hannan (1965) or Diaconis (1988). A brief discussion of essential representation theory appears in Appendix A. The prior distribution $f(m)$ over models is then a discrete distribution over invariant subspaces indexed by m. Where multiplicities arise, the non-uniqueness of the irreducible decomposition makes this task less straightforward; see Forster (2009) for details.

For many common structures, the irreducible decompositions are well known. For example, there is a clear connection with the study of invariant normal linear models, such as those considered by Consonni and Dawid (1985), as invariant linear models for a normal mean, and log-linear models for a Poisson mean coincide if the permutation group under consideration is the same. We next consider the two most common situations.

4.1 Univariate Categorical data

We first consider the case where classification of individuals is with respect to a single categorical variable A with n levels, and there is no further structure to the classification. Hence, we can have either Poisson sampling, $c = 0$, or simple multinomial sampling, $c = 1$.

In such an example, particularly if the classification is with respect to a nominal scale variable, it is common to restrict consideration to classes of models which are invariant under any permutation of the labels of A. The group of permutations of n labels is the symmetric group S_n and the natural permutation matrix representation acting on $\log \boldsymbol{\mu} \in \mathcal{R}^n$ consists of all $n!$ $n \times n$ permutation matrices.

It is well known that there are only two non-trivial S_n-invariant subspaces of $\mathcal{R}^n$, namely $\mathbf{1}_n$ and $N(\mathbf{1}_n^T)$. (Here, and henceforth, $\mathbf{1}_n$ denotes the one dimensional vector subspace of $\mathcal{R}^n$, spanned by $\mathbf{1}_n$). Therefore, the four S_n-invariant log-linear models for $\boldsymbol{\mu}$ are $\mathbf{0}_n$, $\mathbf{1}_n$, $N(\mathbf{1}_n^T)$ and $\mathcal{R}^n$, and can be interpreted as all cell means are one, all cell means are equal, log cell means sum to zero, and the saturated model, respectively.

Under multinomial sampling, $C = \mathbf{1}_n^T$ and the two S_n-invariant log-linear models for $\boldsymbol{\theta}$ correspond to vector spaces 0_n and $N(\mathbf{1}_n^T)$, and can be interpreted as all cell probabilities are equal, and the saturated model, respectively.

4.2 Contingency Tables

Next, we consider multivariate categorical data, where individuals are classified by each of k nominal variables, which we denote $1, \ldots, k$, with corresponding number of levels $r_1, \ldots, r_k$. Hence, the models required are invariant to any combination of permutations of levels of any of the variables concerned. The appropriate permutation group is the direct product

$$G = \prod_{i=1}^{k} S_{r_i}.$$

The natural permutation matrix representation of G acting on the cells of the contingency table (with the elements of $\boldsymbol{p}$ and $\boldsymbol{\theta}$ ordered in a suitable lexographic way) is through the permutation matrices

$$\boldsymbol{P}_g = \bigotimes_{i=1}^{k} \boldsymbol{P}_{g_i}.$$

As there are no multiplicities, the decomposition of $\mathcal{R}^n$ into inequivalent irreducible G-invariant subspaces is given uniquely by

$$\mathcal{R}^n \quad = \quad \bigotimes_{i=1}^{k} \mathbf{1}_{r_i} \oplus N(\mathbf{1}_{r_i}^T) \tag{7}$$

The right hand side of (7) is a direct sum of 2^k tensor product spaces, all of which are G-invariant subspaces of $\mathcal{R}^n$. There are therefore 2^k G-invariant irreducible subspaces of $\mathcal{R}^n$, each of which can be represented by a binary k-vector $\boldsymbol{\gamma}$, where $\gamma_i = 0$ if the ith term in the tensor product is $N(\mathbf{1}_{r_i}^T)$ and $\gamma_i = 1$ otherwise. Therefore $\boldsymbol{\gamma} \in \{0,1\}^k = \Delta$ and an alternative way of expressing (7) is

$$\mathcal{R}^n \quad = \quad \bigoplus_{\boldsymbol{\gamma}\in\{0,1\}^k} \bigotimes_{i=1}^{k} \begin{cases} N(\mathbf{1}_{r_i}^T) & \text{if } \gamma_i = 1 \\ \mathbf{1}_{r_i} & \text{if } \gamma_i = 0 \end{cases} \tag{8}$$

For the permutation invariant inner product $\boldsymbol{I}_n$, the orthogonal projection matrices, onto the 2^k irreducible G-invariant subspaces of $\mathcal{R}^n$ take the form

$$\boldsymbol{Q}_{\boldsymbol{\gamma}} \quad = \quad \bigotimes_{i=1}^{k} \gamma_i \left(\boldsymbol{I}_{r_i} - \frac{1}{r_i}\boldsymbol{J}_{r_i}\right) + (1-\gamma_i)\frac{1}{r_i}\boldsymbol{J}_{r_i} \tag{9}$$

The irreducible G-invariant subspaces are immediately familiar. They represent the usual main effects and interaction terms of a standard loglinear interaction model for a multiway contingency table, where $\boldsymbol{Q}_{\boldsymbol{\gamma}} \log \boldsymbol{\mu}$ is the interaction between all variables i for which $\gamma_i = 1$. In this case, the G-invariant log-linear models, each corresponding to a subset m of $\Delta = \{0,1\}^k$ are exactly the class of log-linear interaction models. Diaconis (1988, p168) discusses this for a 2^k table. Knuiman and Speed (1988) present the projection matrices $\boldsymbol{Q}_{\boldsymbol{\gamma}}$ for a $2\times3\times4$ contingency table. In practice, for reasons of interpretability or computation, consideration is often restricted to those log-linear interaction models which are hierarchical, graphical or decomposable. See Darroch, Lauritzen and Speed (1980) for details. McCullagh (2000) considers further invariance restrictions, under selection of levels of classifying variables, where the invariant models are the hierarchical models.

Under simple multinomial sampling, $\boldsymbol{C} = \mathbf{1}_n^T$, the model is parameterised by $\boldsymbol{\theta} \in N(\mathbf{1}_n^T)$. All G-invariant subspaces of $\mathcal{R}^n$ in (7), except $\mathbf{1}_n$ are G-invariant subspaces of $N(\mathbf{1}_n^T)$. Product multinomial sampling for multiway contingency tables typically involves the totals for the marginal cross-classification of some subset L of the k classifying variables being fixed in advance. Then

$$\boldsymbol{C}_L \quad = \quad \bigotimes_{i=1}^{k} \lambda_i \boldsymbol{I}_{r_i} + (1-\lambda_i)\mathbf{1}_{r_i}^T \tag{10}$$

where the indicator $\lambda_i = 1$ if variable i is in L, and 0 otherwise. If $L = \emptyset$, then we have the simple multinomial constraint. It can be seen that any component of the sum in (8) is in $N(\boldsymbol{C}_L)$ if there exists a variable i for which $\gamma_i = 1$ and $\lambda_i = 0$. Hence the G-invariant decomposition of $N(\boldsymbol{C}_L)$ is of exactly the same form as (8), but with $\{0,1\}^k$ replaced by $\Delta(\boldsymbol{C}_L) = \{\boldsymbol{\gamma} \in \{0,1\}^k : (\mathbf{1} - \boldsymbol{\lambda})^T \boldsymbol{\gamma} > 0\}$. Therefore G-invariant multinomial log-linear models for contingency tables with fixed margins defined by the variable set L, correspond to $m \subseteq \Delta(\boldsymbol{C}_L)$. The terms appearing in $\mathcal{R}^n$, but not in $N(\boldsymbol{C}_L)$ are the intercept and any main effects or interactions involving only variables in the fixed marginal cross-classification. These are also the terms which must be included in a Poisson likelihood analysis, to ensure valid inferences under product multinomial sampling.

5 Prior Distributions for Model Parameters

Each log-linear model m requires a prior distribution for its model parameters $\boldsymbol{\beta}^m$. Again, the prior distribution for the cell probabilities should be constructed in a way which respects invariance considerations. Furthermore, by restricting prior distributions to those which are invariant under certain permutations of the cell labels, the burden of prior specification may be substantially reduced.

Here, we will restrict attention to invariant means and covariance structures, required to specify a multivariate normal prior for $\boldsymbol{\beta}^m$. Rather than considering an explicit parameterisation for an invariant model, for the moment we will focus on the prior mean and covariance for $\log \boldsymbol{\mu}$ under the saturated model. Suppose that the prior mean for $\log \boldsymbol{\mu}$ is $\boldsymbol{\alpha}$ and that the prior variance matrix is $\boldsymbol{\Sigma}$. It is required to find $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ so that the prior distribution is invariant under any permutation $g \in G$. Therefore, $\boldsymbol{\alpha} = \boldsymbol{P}_g \boldsymbol{\alpha}$ and $\boldsymbol{\Sigma} = \boldsymbol{P}_g \boldsymbol{\Sigma} \boldsymbol{P}_g^T$ for all $g \in G$. This implies that any G-invariant mean $\boldsymbol{\alpha}$ is itself G-invariant and hence must lie in the direct sum of all the G-invariant subspaces of $\mathcal{R}^n$ which correspond to the trivial representation (subspaces containing those $\log \boldsymbol{\mu}$ for which $\boldsymbol{P}_g \log \boldsymbol{\mu} = \log \boldsymbol{\mu}$ for any $g \in G$) and which are components of the G-invariant model under consideration.

Determining covariance matrices $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma} = \boldsymbol{P}_g \boldsymbol{\Sigma} \boldsymbol{P}_g^T$ for all $g \in G$ is equivalent to determining $\boldsymbol{\Sigma}$ for which $\boldsymbol{\Sigma} \boldsymbol{P}_g = \boldsymbol{P}_g \boldsymbol{\Sigma}$, as permutation matrices are orthogonal. In other words, we require a set of variance matrices which commute with every matrix $\boldsymbol{P}_g$ of the permutation matrix representation $\boldsymbol{P}_G$ of G. The set of all such matrices form an algebra, which is referred to as the commuting algebra or the commutant algebra of ρ; see Ledermann (1977; 1.8) for details. We are concerned with those elements of the commutant algebra which are symmetric and non-negative definite, and which may therefore be considered as covariance matrices. McLaren (1963) studies the set comprising the symmetric members of the commutant algebra of $\boldsymbol{P}_G$.

Suppose that $\boldsymbol{\Sigma}$ is a member of the commutant algebra of $\boldsymbol{P}_G$, and that $\mathcal{R}^n = \bigoplus_i V_i$ is the canonical decomposition of $\mathcal{R}^n$ into G-invariant subspaces. As before, i indexes inequivalent irreducible representations. Now suppose that $\boldsymbol{T}$ is a unitary matrix with columns composed of elements of a unitary basis for $\mathcal{R}^n$, in such a way that there exists a subset of $\dim(V_i)$ columns of $\boldsymbol{T}$ which form a unitary basis for V_i, and that these columns appear consecutively in $\boldsymbol{T}$, as submatrix $\boldsymbol{T}_i$. It is always possible to construct such a basis. See, for example, Fässler and Stiefel (1992 pp.115–17). [As we are primarily concerned with representations which can be expressed over the real field (all representations of symmetric groups or their direct products), then $\boldsymbol{T}$ is an orthogonal matrix whose columns form an orthonormal basis for $\mathcal{R}^n$.] Then, provided that the submatrices $\boldsymbol{T}_i$ are chosen appropriately if multiplicity $e_i > 1$, the commutant algebra of $\boldsymbol{P}_G$ consists of matrices which can be written as

$$\boldsymbol{\Sigma} \;=\; \boldsymbol{T}\left[\bigoplus_{i=1}^{l} \boldsymbol{\Sigma}^i \otimes \boldsymbol{I}_{d_i}\right]\boldsymbol{T}^{-1} \tag{11}$$

where l is the number of distinct irreducible components of $\boldsymbol{P}_G$, $\boldsymbol{\Sigma}^i$ is an arbitrary $e_i \times e_i$ matrix and d_i is the dimension of each irreducible subspace corresponding to ρ_i; see, for example, Ledermann (1977; pp.29–31) or McLaren (1963). Furthermore, McLaren (1963) shows that for $\boldsymbol{\Sigma}$ to be real and symmetric, we now require each $\boldsymbol{\Sigma}^i$, $i = 1, \ldots, l$, to be real and symmetric. It is straightforward to see, using (11), that non-negative definiteness of each $\boldsymbol{\Sigma}^i$ is a necessary and sufficient condition for $\boldsymbol{\Sigma}$ to be non-negative definite.

Henceforth, we restrict consideration to examples, such as those considered in Sections 4.1 and 4.2, where no multiplicity is greater than one. Then $\boldsymbol{\Sigma}^i = \sigma_i$ is a scalar and we can write

$$\boldsymbol{\Sigma} \;=\; \sum_{i=1}^{l} \sigma_i \boldsymbol{T}_i {\boldsymbol{T}_i}^T. \tag{12}$$

The terms in the summation of (12) are simply non-negative multiples of the projection matrices onto the corresponding V_i, with respect to the invariant inner product I_n. For the non-saturated G-invariant log-linear model $\log \boldsymbol{\mu} \in \bigoplus_{i \in m} V_i$, the prior variance is obtained by setting $\sigma_i = 0$ in (12) unless $i \in m$. The columns of the matrices $\boldsymbol{T}_i$, $i \in m$ lead to a parameterisation of model m through $\boldsymbol{\beta}_i^m = {\boldsymbol{T}_i}^T \log \boldsymbol{\mu}$, $i \in m$. Then $\log \boldsymbol{\mu} = \boldsymbol{T}\boldsymbol{\beta}^m = \sum_{i \in m} \boldsymbol{T}_i \boldsymbol{\beta}_i^m$ and a G-invariant prior distribution for $\boldsymbol{\beta}^m = \{\boldsymbol{\beta}_i^m, i \in m\}$ has covariance matrix

$$\boldsymbol{\Sigma} \;=\; \bigoplus_{i \in m} \sigma_i \boldsymbol{I}_{d_i}. \tag{13}$$

Hence, with this orthonormal parameterisation, G-invariance requires the log-linear model parameters to be uncorrelated. Where multiplicities exist, this constraint may be relaxed; see Forster (2009) for details.

Recall that a (product) multinomial log-linear model, $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} \in R(\boldsymbol{X}) \subseteq N(\boldsymbol{C})$, can be analysed as the Poisson log-linear model $\log \boldsymbol{\mu} \in R(\boldsymbol{X}) \oplus R(\boldsymbol{C}^T)$, subject to (6), where $\boldsymbol{\phi}$ are the 'additional' parameters describing $\log \boldsymbol{\mu} \in R(\boldsymbol{C}^T)$. As mentioned in Section 4, we only consider permutations which do not alter the strata (rows of $\boldsymbol{C}$). Hence, $R(\boldsymbol{C}^T)$ is a G-invariant subspace of $\mathcal{R}^n$ and can be expressed as $R(\boldsymbol{C}^T) = \bigoplus_{i\in\Delta\setminus\Delta(\boldsymbol{C})} V_i$. Then, any Poisson model $\bigoplus_{i\in m} V_i$ where $m \supseteq \Delta \setminus \Delta(\boldsymbol{C})$ can be used to provide marginal inferences for the corresponding multinomial log-linear model $\bigoplus_{i\in m\cap\Delta(\boldsymbol{C})} V_i$ provided that the prior distribution for the model parameters satisfies

$$f(\{\boldsymbol{\beta}_i, i \in m\}) = f(\{\boldsymbol{\beta}_i, i \in m \cap \Delta(\boldsymbol{C})\}). \tag{14}$$

For a multivariate normal prior, this is readily achieved by setting appropriate $1/\sigma_i$ to be zero in the prior precision matrix $S = \bigoplus_{i\in m} \frac{1}{\sigma_i}\boldsymbol{I}_{d_i}$ for the parameters of the Poisson model. More concrete examples follow below.

5.1 Prior distributions for Univariate Categorical data

Recall from Section 4.1 that the decomposition of $\mathcal{R}^n$ into irreducible S_n-invariant subspaces is $\mathcal{R}^n = \mathbf{1}_n \oplus N(\mathbf{1}_n^T)$. The invariant subspace $\mathbf{1}_n$ corresponds to the trivial representation, so for models where this component is present, a prior mean $\boldsymbol{\alpha} \propto \mathbf{1}_n$ is permitted. As $\mathbf{1}_n$ and $N(\mathbf{1}_n^T)$ correspond to inequivalent representations, any S_n-invariant covariance matrix for $\log \boldsymbol{\mu}$ must be of the form

$$\boldsymbol{\Sigma} \;=\; \frac{\sigma_1}{n}\boldsymbol{J} + \sigma_2\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right).$$

Here σ_1 controls the prior uncertainty about the overall size of the cell means, while σ_2 reflects the strength of prior belief in equal cell probabilities, with the limiting value $\sigma_2 = 0$ corresponding to the null model of a common cell mean. Marginal inferences under a Poisson model will be valid under simple multinomial sampling, where $\boldsymbol{C} = \mathbf{1}^T$ (and hence $\boldsymbol{C}^T = \mathbf{1}$) provided that we set $1/\sigma_1 = 0$ in the precision $\boldsymbol{S} = \frac{1}{\sigma_1 n}\boldsymbol{J} + \frac{1}{\sigma_2}\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)$ of a multivariate normal prior for $\log \boldsymbol{\mu}$. Alternatively, as discussed in Section 2, independent gamma priors for the cell means will suffice (with common parameters if permutation invariance is required).

5.2 Contingency Tables

For those $r_1 \times \cdots \times r_k$ tables considered in Section 4.2, where the categorical variables are considered to be nominal, and $G = \prod_{i=1}^k S_{r_i}$, the G-invariant decomposition is given by (8). Again, the invariant subspace $\mathbf{1}_n = \bigotimes_i \mathbf{1}_{r_i}$ corresponds to the trivial representation, and for models where this component is present, a prior mean $\boldsymbol{\alpha} \propto \mathbf{1}$ is permitted. No irreducible representation occurs with multiplicity greater than one, so we index each invariant subspace, and corresponding prior variance term by by its corresponding interaction label $\boldsymbol{\gamma} \in \{0,1\}^k$. Hence, a G-invariant covariance matrix must take the form

$$\boldsymbol{\Sigma} \;=\; \sum_{\boldsymbol{\gamma}\in\{0,1\}^k} \sigma_{\boldsymbol{\gamma}} \boldsymbol{Q}_{\boldsymbol{\gamma}}$$

where the $\boldsymbol{Q}_{\boldsymbol{\gamma}}$, given by (9), are the projection matrices onto the irreducible G-invariant subspaces and $\sigma_{\boldsymbol{\gamma}} = 0$ for any term (subspace) not included in the model under consideration. Hence the prior requires specification of a single dispersion parameter for every term present in the model. For any parameterisation of a log-linear interaction model where the columns of the model matrix $\boldsymbol{X}$ are orthonormal, the parameters must be *a priori* uncorrelated, and parameters corresponding to the same main effect or interaction term must have common variance if the prior distribution is to be invariant under G.

While it is not necessary to construct the prior with respect to an orthonormal parameterisation, the resulting marginal prior distribution for such a parameterisation must have these (independence and common variance) properties unless prior information suggests that invariance under G is not appropriate. Different parameterisations are linearly related, so this can be checked. An orthonormal model matrix can easily be constructed, for example, by a Gram-Schmidt procedure using columns of $\boldsymbol{Q}_{\boldsymbol{\gamma}}$ for each $\boldsymbol{\gamma} \in m$. Alternatively the standard parameterisation of a log-linear interaction model using ‘sum-to zero’ constraints on model parameters produces a model matrix where columns corresponding to different model terms $\boldsymbol{\gamma}$ are naturally orthogonal, although parameters corresponding to the same model term are not, but can easily be made so.

For product multinomial models where the totals for the marginal cross-classification of some subset L of the k classifying variables are fixed in advance, $\boldsymbol{\theta} \in N(\boldsymbol{C}_L)$ where $\boldsymbol{C}_L$ is given by (10). Furthermore $R(\boldsymbol{C}_L^T)$ is given by a direct sum of exactly the same form as (8), but with $\{0,1\}^k$ replaced by $\Delta \setminus \Delta(\boldsymbol{C}_L) = \{\boldsymbol{\gamma} \in \{0,1\}^k : (\mathbf{1} - \boldsymbol{\lambda})^T\boldsymbol{\gamma} = 0\}$. Hence the Poisson model $\boldsymbol{\gamma} \in m \supseteq \Delta \setminus \Delta(\boldsymbol{C}_L)$ can be used to provide valid marginal inferences for the multinomial model $m \cap \Delta(\boldsymbol{C}_L)$ if the prior for the log-linear parameters satisfies (14). Hence for a multivariate normal prior for $\boldsymbol{\beta}$ in the Poisson model, we require $1/\sigma_{\boldsymbol{\gamma}} = 0$ in the prior precision matrices for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, for $\boldsymbol{\gamma} \in \Delta \setminus \Delta(\boldsymbol{C}_L)$, in other words for the parameters corresponding to the ‘intercept’ and all main effects and interactions involving variables in the fixed margin L only.

6 The Posterior Distribution

As the prior for at least some of the parameters of a Poisson model may be improper, we need to consider whether the resulting posterior will be proper. Consider an arbitrary Poisson log-linear model $\log \boldsymbol{\mu} = \boldsymbol{X\beta}$. First, we derive conditions for the posterior distribution for $\boldsymbol{\beta}$ resulting from an improper uniform prior to be proper. In fact, the following is sufficient for any prior for $\boldsymbol{\beta}$ which has bounded density over $\mathcal{R}^p$. We use the fact that a log-linear model is a generalised linear model with canonical link, and hence the likelihood is a log-concave function of $\boldsymbol{\beta}$.

Theorem

A necessary and sufficient condition for a log-concave function to have a finite integral is that it achieves its maximum in the interior of the parameter space. In other words the maximum likelihood estimate for $\boldsymbol{\beta}$ must be finite.

Proof

To prove suffiency, we first note that for any log-concave function $g(\boldsymbol{\beta})$ and any $r > 0$, there exists finite positive numbers a and b such that $g(\boldsymbol{\beta}) < a\exp(-b|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|)$ for all $\boldsymbol{\beta} \notin R = \{|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}| < r^2\}$, where $\hat{\boldsymbol{\beta}}$ is the mode of g. Then,

$$\int_{\mathcal{R}^p} g(\boldsymbol{\beta})\, d\boldsymbol{\beta} \;\; < \;\; \int_R g(\boldsymbol{\beta})\, d\boldsymbol{\beta} \; + \; a\int_{\mathcal{R}^p} \exp(-b|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}|)\, d\boldsymbol{\beta},$$

and both the integrals on the right hand side are finite. (The first is a bounded function, over a finite region, the second is equal to $2(\pi/b^2)^{p/2}\Gamma(p)/\Gamma(p/2)$.) To observe the necessity of a finite maximum likelihood estimate, note that for a log-concave density function the surfaces of equal density are concave. If the mle is infinite, they each divide $\mathcal{R}^p$ into two regions of infinite volume. In one of these regions, the density is bounded above zero and hence its integral is unbounded.

Now for the Poisson log-linear model, the log-likelihood is

$$L(\boldsymbol{\beta}) \;\; = \;\; -\sum_{i=1}^{n} \exp[\boldsymbol{X\beta}]_i \; + \; \boldsymbol{y}^T\boldsymbol{X\beta}$$

Consider a parameterisation $\boldsymbol{\beta} = |\boldsymbol{\beta}|\boldsymbol{\beta}^u$, where $|\boldsymbol{\beta}^u| = 1$. Then

$$L(|\boldsymbol{\beta}|, \boldsymbol{\beta}^u) \;\; = \;\; -\sum_{i=1}^{n} \exp[|\boldsymbol{\beta}|\boldsymbol{X\beta}^u]_i \; + \; |\boldsymbol{\beta}|\boldsymbol{y}^T\boldsymbol{X\beta}^u. \tag{15}$$

For $\boldsymbol{\beta}$ to have a finite mle, we require $L \to -\infty$ as $|\boldsymbol{\beta}| \to \infty$, for every $\boldsymbol{\beta}^u$. For a given $\boldsymbol{\beta}^u$, let $\delta = \max_i \exp[\boldsymbol{X}\boldsymbol{\beta}^u]_i$ and denote the number of $\exp[\boldsymbol{X}\boldsymbol{\beta}^u]_i$ which attain δ by d. If $\delta > 0$, then $L = -\exp(|\boldsymbol{\beta}|\delta)[d + o(1)]$, so clearly $L \to -\infty$ as $|\boldsymbol{\beta}| \to \infty$. If $\delta \leq 0$, then $L = |\boldsymbol{\beta}|\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}^u - O(1)$. Therefore, a necessary and sufficient condition for a finite mle is $\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} < 0$ for all $\boldsymbol{\beta} \neq \boldsymbol{0}$ such that $\boldsymbol{X}\boldsymbol{\beta} \leq \boldsymbol{0}$ where the second inequality is in every component. This condition was first proved by Haberman (1974, Theorem 2.3). Clearly, for any model which permits $\boldsymbol{X}\boldsymbol{\beta} < \boldsymbol{0}$, (all models with $\boldsymbol{1}$ as a column of $\boldsymbol{X}$) we will require $\sum_{i=1}^n y_i > 0$; at least one positive cell count. Further constraints on $\boldsymbol{y}$ for the posterior to be proper arise by considering other possible linear predictors $\boldsymbol{X}\boldsymbol{\beta}$ which are non-positive with at least one zero component. Glonek, Darroch and Speed (1988) describe the implications of this result for hierarchical log-linear models.

We now generalise this result to posterior distributions resulting from a particular improper prior distribution, the multivariate normal

$$f(\boldsymbol{\beta}) \quad \propto \quad \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\alpha})^T\boldsymbol{S}(\boldsymbol{\beta}-\boldsymbol{\alpha})\right\} \qquad (16)$$

where $\boldsymbol{S}$ is non-negative definite, of rank $q < p$. As the posterior density will be log-concave, a necessary and sufficient condition for it to be proper, as before, is that it achieves its maximum in the interior of the parameter space, and hence that $L + \log f \to -\infty$ as $|\boldsymbol{\beta}| \to \infty$, where L and f are given by (15) and (16) respectively. As $\log f$ is quadratic in $|\boldsymbol{\beta}|$, this will be the case for all $\boldsymbol{\beta} \notin N(\boldsymbol{S})$, and hence the condition becomes $\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} < 0$ for all $\boldsymbol{\beta} \in N(\boldsymbol{S})\setminus\{\boldsymbol{0}\}$ such that $\boldsymbol{X}\boldsymbol{\beta} \leq \boldsymbol{0}$. For the Poisson models considered in Section 3, if the prior is proper for $\boldsymbol{\beta}$, we need consider only $\boldsymbol{\phi}$, and hence situations in which $\boldsymbol{C}^T\boldsymbol{\phi} \leq \boldsymbol{0}$. It is straightforward to see that $\boldsymbol{C}\boldsymbol{y} > \boldsymbol{0}$ is a necessary and sufficient condition for the posterior to be proper. In other words, each of the prespecified group totals, N_i, must be positive.

7 Examples

We now present two small examples to illustrate some of the ideas presented in the paper.

7.1 Example 1

For illustration, we present a possible Bayesian analysis of the product binomial example presented by Lang (1996). Here, $c = 2$, $n_1 = n_2 = 2$, $\boldsymbol{y} = (30, 20, 60, 15)^T$, $\boldsymbol{N} = (50, 75)^T$, $L = \{1\}$, $\lambda = (1, 0)^T$ and

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

so $p_1 + p_2 = p_3 + p_4 = 1$. The centred logits are given by $\theta_1 = -\theta_2 = \frac{1}{2}(\log p_1 - \log p_2)$ and $\theta_3 = -\theta_4 = \frac{1}{2}(\log p_3 - \log p_4)$. The saturated Poisson log-linear model may be expressed as $\boldsymbol{\theta} = \boldsymbol{T\beta}$ where

$$\boldsymbol{T} \;=\; \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Each of the columns of $\boldsymbol{T} = (\boldsymbol{T}_{00}, \boldsymbol{T}_{10}, \boldsymbol{T}_{01}, \boldsymbol{T}_{11})$ is an orthonormal basis for an inequivalent one-dimensional invariant subspace for $\log \boldsymbol{\mu}$. Columns $\boldsymbol{T}_{01}$ and $\boldsymbol{T}_{11}$) span $N(\boldsymbol{C})$ and hence form orthonormal bases for inequivalent one-dimensional invariant subspaces for $\boldsymbol{\theta}$. Possible permutation invariant models have model matrices whose columns are a subset of those in $\boldsymbol{T}$. In this example, we consider two possible models: the saturated model, and the model with no interaction (final column absent). We consider these models to be a priori equally probable.

We construct a multivariate normal prior for $\boldsymbol{\beta}$ which respects invariance under permutation of the row or column labels. Hence, the model parameters $\boldsymbol{\beta} = (\beta_{00}, \beta_{10}, \beta_{01}, \beta_{11})$ must be independent *a priori*. The prior mean for $(\beta_{10}, \beta_{01}, \beta_{11})$ must be $\mathbf{0}$, but β_{00} is allowed a non-zero mean in Poisson models where it is present. However, for valid multinomial inferences to be obtained from Poisson models, we need to set the prior precision for β_{00} and β_{10} to zero. Hence, the resulting improper prior distribution for the saturated Poisson model is

$$f(\beta_{00}, \beta_{10}, \beta_{01}, \beta_{11}) \;\propto\; \exp\left\{-\frac{1}{2\sigma_{01}}\beta_{01}^2 - \frac{1}{2\sigma_{11}}\beta_{11}^2\right\}. \tag{17}$$

The posterior will necessarily be proper, from the results of Section 6, as $\boldsymbol{Cy} > \mathbf{0}$ by design. The prior for (β_{01}, β_{11}) for the saturated multinomial model is given by the right hand side of (17). For the no interaction model, β_{11} is not present, and the corresponding term vanishes from the prior.

In the following analysis, we set $\sigma_{01} = \pi^2/2 \approx 4.935$, and give σ_{11} the same value in the saturated model. Hence in the saturated model, the prior for the logits of the independent product binomial probabilities has the same mean and variance as the corresponding independent Jeffreys priors. For inference, we focus on the posterior marginal densities of β_{01} (common log odds) for the no interaction model and β_{11} (0.5 times log odds ratio) for the saturated model. The latter also enables us to calculate the Bayes factor for comparing the models, using the Savage-Dickey density ratio of the prior and posterior densities of β_{11} at 0. The posterior densities are calculated using Laplace's method, although the Gibbs sampler is also extremely convenient for log-linear models, which have log-concave posterior densities.

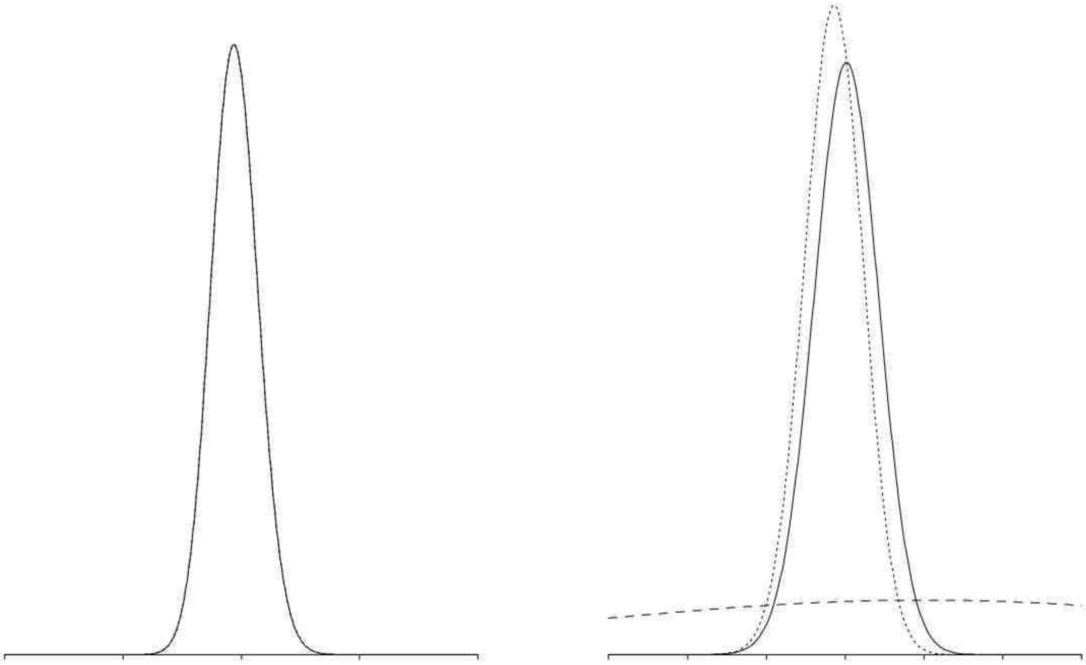

Figure 1: Posterior marginal densities for β_{01} in the no interaction model (a) and β_{11} in the saturated model (b). Panel (b) also displays the prior (dashed line). Both panels also display the posterior marginal density derived from the Poisson model with $\sigma_{10} = 10^{-3}$ (dotted line), although in (a) this is indistinguishable from the true multinomial posterior density.

Figure 1 displays the posterior marginal densities for β_{01} in the no interaction model (a) and β_{11} in the saturated model (b). The Bayes factor in favour of the saturated model is 1.74, calculated directly using Laplace's method or using the Savage-Dickey density ratio. For illustration, the plots also contain 'incorrect' marginal densities derived from a Poisson analysis where σ_{10} is finite. In particular, the inference concerning model comparison is potentially misleading, as the Bayes factor in favour of the saturated model increases to 10.94.

7.2 Example 2

Here we consider an example of univariate categorical data where it is natural to consider a permutation group other than the symmetric group S_n in constructing the prior. The top line of Table 1 is taken from Santner and Duffy (1989, p.95) and represents cases of Acute Lymphatic Leukaemia recorded in the British Cancer Registry from 1946–60, classified by month of entry. Santner and Duffy find that the model of uniform monthly rate of entry is a poor fit to these data.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Cases	39	58	51	56	36	48	33	38	40	34	30	44
Estimates	42.5	48.5	45.8	49.1	43.4	44.3	38.4	39.8	35.9	39.4	37.7	43.3

Table 1: Cases of Acute Lymphatic Leukaemia recorded in the British Cancer Registry from 1946–60, classified by month of entry, together with the corresponding Bayes estimates (posterior expected cell means) obtained by 'model averaging' over Fourier regression models.

As the data are classified by a variable which is cyclic, it seems sensible here to consider models and prior distributions which are invariant under C_n, the cyclic permutations of the category labels (months). A little care is required with the representation theory, as real and complex representations do not coincide (unlike symmetric groups). As $n = 12$ is even then $\mathcal{R}^n$ has two one-dimensional C_n-invariant irreducible subspace spanned by $\boldsymbol{c}^0 = \mathbf{1}$ and the alternating vector $\boldsymbol{c}^{n/2} = (-1, 1, -1, 1, \ldots, -1, 1)^T$ respectively. Note that $c_k^{n/2} = \cos k\pi$. The remaining irreducible C_n invariant subspaces are two dimensional and are spanned by $\{\boldsymbol{c}^{l1}, \boldsymbol{c}^{l2}\}$ for $l = 1, \ldots, n/2 - 1$ where $(\boldsymbol{c}_k^{l1}, \boldsymbol{c}_k^{l2}) = (\cos 2kl\pi/n, \sin 2kl\pi/n)$. These spaces represent cosine curves with period n/l (frequency l) for $l = 1, \ldots, (n-1)/2$. For n odd, the decomposition is similar, but there is no space equivalent to $\boldsymbol{c}^{n/2}$. The resulting models are log-linear Fourier regression models with an evenly spaced covariate. For a cyclic factor with 12 levels, there are five invariant subspaces of dimension 2, representing cosine curves of frequency $1, \ldots, 5$ and two of dimension 1, representing a constant effect and a cosine curve of frequency 6, respectively. This results in a total of 128 possible Poisson log-linear models. In our analysis, we will assume that the 'intercept' ($l = 0$) is present in all models under consideration.

All the basis vectors derived above are orthogonal, and can be normalised to construct the orthonormal model matrix $\boldsymbol{T}$ for any given invariant model. A C_n-invariant prior requires a zero mean for any $\boldsymbol{\beta}_l$ other than $\boldsymbol{\beta}_0$ (corresponds to trivial representation). In the current example, we choose to set $E(\boldsymbol{\beta}_0) = \mathbf{0}$. Each irreducible component has a single prior variance parameter so, for example, $Var(\beta_{l1}, \beta_{l2}) = \sigma_l I_2$. The resulting C_n-invariant covariance matrix for $\log \boldsymbol{\mu}$ may be expressed as $\boldsymbol{\Sigma}$ where

$$\boldsymbol{\Sigma}_{jk} \;=\; \sum_{l=1}^{[n/2]} \frac{a_l \sigma_l}{n} \cos\left(2\pi l(j-k)/n\right) \qquad i, j = 1, \ldots, n \tag{18}$$

where $a_l = 1$ if $l = 0$ or $l = n/2$ and $a_l = 2$ otherwise (see also, Dawid and Consonni, 1985). The $[n/2] + 1$ components of this matrix are the projection matrices onto the $[n/2] + 1$ real invariant subspaces, and hence non-negative definiteness is ensured if and only if all σ_l are non-negative.

We present two possible Bayesian analyses of the data in Table 1. The first is based on calculating posterior model probabilities using the Bayesian Information Criterion (BIC). This is a crude, but simple, way of calculating posterior model probabilities using model deviances, which does not require specification of a prior distribution for the model parameters. See Kass and Raftery (1995) for details. We also present an alternative fully Bayesian analysis, specifying an invariant normal prior for the parameters of each model. For $\boldsymbol{\beta}_0$ we set $\sigma_0 \to \infty$. For all other $\boldsymbol{\beta}_l$ present in a model, we choose a proper, but diffuse, prior by specifying prior variances $\sigma_l = \psi'(\lambda), i = 1, \ldots, 6$. Then, the prior mean and variance for the corresponding multinomial $\boldsymbol{\theta}$ are the same as for a symmetric Dirichlet prior for multinomial cell probabilities, with all parameters equal to λ. Here we use $\lambda = \frac{1}{2}$ (Jeffreys' prior), in which case $\psi'(\lambda) = \pi^2/2 \approx 4.935$.

Two sets of posterior model probabilities are presented in Table 2. For the full Bayesian analysis thay have been calculated using Laplace's method. It can be seen that qualitatively the results are very similar, with exactly the same four models having non-negligible ($> 10^{-2}$) posterior probability, accounting for over 98% of total probability. These models are the null model, and the models with frequency 1, frequency 6 and frequencies 1 and 6 together. The frequency one term represents a yearly cycle of admissions, and the frequency 6 term a bimonthly fluctuation, which is more difficult to interpret. The model-averaged estimated cell means, calculated using Laplace's method and presented in the second row of Table 1, are largely based on these four models.

Appendix A Group Representations

The natural representation, ρ (which we also denote P_G), of the action of G on $\mathcal{R}^n$ maps $g \in G$ to $\boldsymbol{P}_g$. When ρ is restricted to a G-invariant subspace V_i of $\mathcal{R}^n$, then the resulting subrepresentation, ρ_i, maps $g \in G$ to the corresponding linear transformation in V_i. Therefore invariant subspaces of $\mathcal{R}^n$ correspond to subrepresentations of ρ.

Model	Posterior probability		df
	BIC	Bayes	
null	0.4215	0.2323	11
1	0.2445	0.3209	9
6	0.2006	0.1708	10
1+6	0.1163	0.2360	8

Table 2: Posterior model probabilities for Table 1, for models with probabilities greater than 10^{-2}, calculated using BIC, and using a fully Bayesian approach. A model is denoted by the frequencies of the cosine functions present.

An irreducible representation of G is one which has no non-trivial subrepresentation, and every representation is a direct sum of irreducible subrepresentations. In other words, $\mathcal{R}^n$ can be decomposed as a direct sum of minimal G-invariant subspaces $\mathcal{R}^n = \bigoplus_i V_i$. However, this decomposition is not necessarily unique. If the action of G on V_j is isomorphic to the action of G on V_k then the corresponding subrepresentations ρ_j and ρ_k are said to be equivalent. There are then an infinite number of ways of decomposing $V_j \oplus V_k$ into two irreducible invariant subspaces. The number of times an equivalent representation apppears in an irreducible decomposition is called the multiplicity of the representation in the decomposition. The canonical decomposition is $\mathcal{R}^n = \bigoplus_i V_i$ where i indexes inequivalent irreducible representations and V_i is the direct sum of the e_i invariant subspaces corresponding to ρ_i, where e_i is the multiplicity of ρ_i in ρ. The canonical decomposition is unique, but is reducible if any multiplicity is greater than 1.

In all the examples presented in this paper, all multiplicities are one, the unique canonical decomposition $\mathcal{R}^n = \bigoplus_i V_i$ is irreducible and any G-invariant Poisson log-linear model can be expressed as $\bigoplus_{i\in m} V_i$ where m is any subset of Δ, the index set for the irreducible components. Hence there are $2^{|\Delta|}$ possible G-invariant models, each corresponding to a particular $m \subseteq \Delta$. An invariant multinomial log-linear model corresponds to any $\bigoplus_{i\in m} V_i \subseteq N(\boldsymbol{C})$.

References

Aitchison J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Baker, S. G. (1994). The multinomial-Poisson transformation. *The Statistician*, **43**, 495–504.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *The Journal of* *the Royal Statistical Society*, B**25**, 220–233.

Consonni G. and Dawid A. P. (1985). Decomposition and Bayesian analysis of invariant normal linear models. *Linear Algebra and its Applications*, **70**, 21–49.

Darroch J. N., Lauritzen S. L. and Speed T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics* **8**, 522–539.

Diaconis P. (1988). *Group Representations in Probability and Statistics.* Hayward, California: Institute of Mathematical Statistics.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *The Journal of the Royal Statistical Society*, B**31**, 195–233.

Fässler and Stiefel (1992). *Group Theoretical Methods and their Applications.* Boston: Birkhäuser.

Forster, J. J. (2009). Bayesian inference for square contingency tables
`http://www.soton.ac.uk/∼jjf/Papers/Square.pdf`.

Glonek, G. F. V., Darroch, J. N. and Speed, T. P. (1988). On the existence of maximum likelihood estimators for hierarchical log-linear models. *Scandinavian Journal of Statistics* **15**, 187–193.

Gûnel E. and Dickey J. (1974). Bayes factors for independence in contingency tables. *Biometrika* **61**, 545–557.

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

Hannan, E. J. (1965). Group representations and applied probability. *Journal of Applied Probability* **2**, 1–68.

James, G. and Liebeck, M. (1993). *Representations and Characters of Groups.* Cambridge: Cambridge University Press.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90**, 773–795.

King, R. and Brooks, S. P. (2001). Prior induction in log-linear models for general contingencty table analysis. *Annals of Statistics*, **29**, to appear.

Knuiman, M. W. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**, 1061–1071.

Lang, J. B. (1996). On the comparison of multinomial and Poisson log-linear models. *The*

Journal of the Royal Statistical Society, B**58**, 253–266.

Ledermann, W. (1977). *Introduction to Group Characters.* Cambridge: Cambridge University Press.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, **35**, 1622–1643.

McCullagh, P. (2000). Invariance and factorial models (with discussion). *The Journal of the Royal Statistical Society*, B**62**, 209–256.

McLaren, A. D. (1963). On group representations and invariant stochastic processes. *Proceedings of the Cambridge Philosophical Society*, **59**, 431–450.