TEST
DOI 10.1007/s11749-017-0540-8

ORIGINAL PAPER

# On the correspondence from Bayesian log-linear modelling to logistic regression modelling with $g$-priors

**Michail Papathomas**[1]

Received: 26 April 2016 / Accepted: 26 April 2017
© The Author(s) 2017. This article is an open access publication

**Abstract** Consider a set of categorical variables where at least one of them is binary. The log-linear model that describes the counts in the resulting contingency table implies a specific logistic regression model, with the binary variable as the outcome. Within the Bayesian framework, the $g$-prior and mixtures of $g$-priors are commonly assigned to the parameters of a generalized linear model. We prove that assigning a $g$-prior (or a mixture of $g$-priors) to the parameters of a certain log-linear model designates a $g$-prior (or a mixture of $g$-priors) on the parameters of the corresponding logistic regression. By deriving an asymptotic result, and with numerical illustrations, we demonstrate that when a $g$-prior is adopted, this correspondence extends to the posterior distribution of the model parameters. Thus, it is valid to translate inferences from fitting a log-linear model to inferences within the logistic regression framework, with regard to the presence of main effects and interaction terms.

**Keywords** Categorical variables · Contingency tables · Mixtures of $g$-priors · Prior correspondence · Posterior correspondence

**Mathematics Subject Classification** 62F15

**Electronic supplementary material** The online version of this article (doi:10.1007/s11749-017-0540-8) contains supplementary material, which is available to authorized users.

✉ Michail Papathomas
M.Papathomas@st-andrews.ac.uk

[1] School of Mathematics and Statistics, University of St Andrews, The Observatory, Buchanan Gardens, St Andrews KY16 9LZ, UK

Published online: 18 May 2017

## 1 Introduction

Consider observations $\boldsymbol{v} = \{v_1, \ldots, v_n\}$, parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$, and known quantities or nuisance parameters $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$. Following standard notation, $v_i, i = 1, \ldots, n$, follows a distribution that is a member of the exponential family when its probability function can be written as,

$$f(v_i|\theta_i, \phi_i) = \exp\left\{\frac{w_i}{\phi_i}\left[v_i\theta_i - b(\theta_i)\right] + c(v_i, \phi_i)\right\},$$

where $\boldsymbol{w} = \{w_1, \ldots, w_n\}$ are known weights, and $\phi_i$ is described as the dispersion or scale parameter. With regard to first- and second-order moments, $\mu_i \equiv E(v_i) = b^{'}(\theta_i)$ and $\text{Var}(v_i) = \frac{w_i}{\phi_i} b^{''}(\theta_i)$. The variance function is defined as $V(\mu_i) = b^{''}(\theta_i)$. A generalized linear model relates $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_n\}$ to covariates by setting $\zeta(\boldsymbol{\mu}) = X_d\boldsymbol{\gamma}$, where $\zeta$ denotes the link function, $X_d$ the covariate design matrix, and $\boldsymbol{\gamma}$ a vector of parameters. For a single $\mu_i$, we write $\zeta(\mu_i) = X_{d(i)}\boldsymbol{\gamma}$, where $X_{d(i)}$ denotes the $i$th row of $X_d$. So, $\zeta$ is defined as a vector function $\zeta \equiv \{\zeta_1, \ldots, \zeta_n\}$ with $n$ elements.

Denote with $\mathcal{P}$ a finite set of $P$ categorical variables. Observations from $\mathcal{P}$ can be arranged as counts in a $P$-way contingency table. Denote the cell counts as $n_i, i = 1, \ldots, n_{ll}$. We use the '$ll$' indicator to allude to the log-linear model that will describe these counts. A Poisson distribution is assumed for the counts so that $E(n_i) = \mu_i$. A Poisson log-linear interaction model $\log(\boldsymbol{\mu}) = X_{ll}\boldsymbol{\lambda}$ is a generalized linear model that relates the expected counts to $\mathcal{P}$. Assuming that one of the categorical variables, denoted with $Y$, is binary, a logistic regression can also be fitted with $Y$ as the outcome, and all or some of the remaining $P - 1$ variables as covariates. We write, $\text{logit}(\boldsymbol{p}) = X_{lt}\boldsymbol{\beta}$, $\boldsymbol{p} = (p_1, \ldots, p_{n_{lt}})$, using the '$lt$' indicator for the logistic model. Here, $p_i$ denotes the conditional probability that $Y = 1$ given covariates $X_{lt(i)}$, and $\boldsymbol{\beta}$ is a vector of parameters.

Within the Bayesian framework, a prior distribution $f(\boldsymbol{\gamma})$ is assigned to the parameters of the log-linear or logistic regression model. This can be an informative prior that incorporates prior information on the magnitude of the effect of the different covariates or interactions. Eliciting such a prior distribution is not straightforward, especially for the coefficients of interaction terms (Consonni and Veronese 2008). Typically, lack of information for the parameters of a generalized linear model leads to a relatively flat but proper prior distribution, so that model determination based on Bayes factors is valid (O'Hagan 1995). A popular choice among Bayesian statisticians is the $g$-prior or a mixture of $g$-priors, described in detail in Sect. 2. These are flexible priors designed to carry very little information so that inferences are driven by the observed data. See, for example, Wang and George (2007), Sabanès Bovè and Held (2011), Overstall and King (2014a, b) and Mukhopadhyay and Samantha (2016). This type of prior was first proposed by Zellner (1986) for general linear models. In this context, it is known as Zellner's $g$-prior. Other priors have been proposed, especially for analyses where the focus is on model comparison and variable selection. For example, Jeffreys prior (Liang et al. 2008), the generalized hyper-$g$ prior (Sabanès Bovè and Held 2011), and the expected-posterior priors and power-expected-posterior priors (Fouskakis et al.

2015). Our manuscript concerns the $g$-prior and mixtures of $g$-priors. After data are collected, the prior $f(\boldsymbol{\gamma})$ is updated to the posterior distribution $f(\boldsymbol{\gamma}|\text{Data})$ via the conditional probability formula and Bayes Theorem, so that,

$$f(\boldsymbol{\gamma}|\text{Data}) = \frac{f(\text{Data}|\boldsymbol{\gamma})\, f(\boldsymbol{\gamma})}{f(\text{Data})}.$$

For the prior distributions discussed above, closed-form expressions for the posterior distribution $f(\boldsymbol{\gamma}|\text{Data})$ do not exist. The posterior is typically calculated using Markov chain Monte Carlo stochastic simulation, or Normal approximations (O'Hagan and Forster 2004).

It is known (Agresti 2002) that when $\mathcal{P}$ contains a binary $Y$, a log-linear model $\log(\boldsymbol{\mu}) = X_{ll}\boldsymbol{\lambda}$ implies a specific logistic regression model with parameters $\boldsymbol{\beta}$ defined uniquely by $\boldsymbol{\lambda}$. The logistic regression model for the conditional odds ratios for $Y$ implies an equivalent log-linear model with arbitrary interaction terms between the covariates in the logistic regression, plus arbitrary main effects for these covariates. We provide a simple example to illustrate this result and clarify additional notation. Assume three categorical variables $X$, $Y$, and $Z$, with $Y$ binary. Let $i, j, k$ be integer indices that describe the level of $X$, $Y$, and $Z$, respectively. For instance, as $Y$ is binary, $j = 0, 1$. Consider the log-linear model,

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \tag{M1}$$

where the superscript denotes the main effect or interaction term. The corresponding logistic regression model for the conditional odds ratios for $Y$ is derived as follows,

$$\begin{aligned}\log\left(\frac{P(Y=1|X,Z)}{P(Y=0|X,Z)}\right) &= \log\left(\frac{P(Y=1,X,Z)}{P(Y=0,X,Z)}\right)\\ &= \log(\mu_{i1k}) - \log(\mu_{i0k})\\ &= \lambda_1^Y - \lambda_0^Y + \lambda_{i1}^{XY} - \lambda_{i0}^{XY} + \lambda_{1k}^{YZ} - \lambda_{0k}^{YZ}.\end{aligned}$$

This is a logistic regression with parameters, $\boldsymbol{\beta} = (\beta, \beta_i^X, \beta_k^Z)$, so that, $\beta = \lambda_1^Y - \lambda_0^Y$, $\beta_i^X = \lambda_{i1}^{XY} - \lambda_{i0}^{XY}$, and $\beta_k^Z = \lambda_{1k}^{YZ} - \lambda_{0k}^{YZ}$. Considering identifiability corner point constraints, all elements in $\boldsymbol{\lambda}$ with a zero subscript are set to zero. Then, $\beta = \lambda_1^Y$, $\beta_i^X = \lambda_{i1}^{XY}$ and $\beta_k^Z = \lambda_{1k}^{YZ}$. This scales in a straightforward manner to larger log-linear models. For instance, if (M1) contained the three-way interaction $XYZ$, then the corresponding logistic regression model would contain the $XZ$ interaction, so that, $\beta_{ik}^{XZ} = \lambda_{i1k}^{XYZ} - \lambda_{i0k}^{XYZ}$, and under corner point constraints, $\beta_{ik}^{XZ} = \lambda_{i1k}^{XYZ}$. If a factor does not interact with $Y$ in the log-linear model, then this factor disappears from the corresponding logistic regression model. To demonstrate that the correspondence between log-linear and logistic models is not bijective, it is straightforward to show that, for example, the log-linear model, $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$, implies the same logistic regression as (M1). More generally, the relation between $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ can be described as $\boldsymbol{\beta} = \boldsymbol{T}\boldsymbol{\lambda}$, where $\boldsymbol{T}$ is an incidence matrix (Bapat 2011). In the context of this manuscript, matrix $\boldsymbol{T}$ has one row for each element of $\boldsymbol{\beta}$, and

one column for each element of $\boldsymbol{\lambda}$. The elements of $\boldsymbol{T}$ are zero, except in the case where the element of $\boldsymbol{\beta}$ is defined by the corresponding element of $\boldsymbol{\lambda}$. The number of rows of $\boldsymbol{T}$ cannot be greater than the number of columns. To simplify the analysis and notation, for the remainder of this manuscript we consider models specified under corner point constraints. Then, every logistic regression model parameter is defined uniquely by the corresponding log-linear model parameter, and the correspondence from a log-linear to a logistic regression model is direct.

The contribution of our manuscript is twofold. First, Theorem 1 states that assigning to $\boldsymbol{\lambda}$ the $g$-prior that is specific to log-linear modelling implies the $g$-prior specific to logistic modelling on the parameters $\boldsymbol{\beta}$ of the corresponding logistic regression. The log-linear model has to be the largest model that corresponds to the logistic regression, i.e. the model that contains all possible interaction terms between the categorical factors in $\mathcal{P}\backslash\{Y\}$. Second, under the reasonable assumption that an investigator who chooses a $g$-prior for $\boldsymbol{\lambda}$ would also choose a $g$-prior for $\boldsymbol{\beta}$ if they were to fit a logistic regression directly, inferences on the parameters of a log-linear model translate to inferences on the parameters of the corresponding logistic regression. Closed-form expressions for the posterior distributions do not exist. Wang and George (2007) utilize the Laplace approximation for generalized linear models, focusing on the approximation of the marginal likelihood for the purpose of variable selection. Theorem 2 shows that, asymptotically, the matching between the prior distributions of the corresponding parameters extends to the posterior distributions. It is then demonstrated by numerical illustrations that the presence or absence of interaction terms in the log-linear model can inform on the relation between the binary $Y$ and the other variables as described by logistic regression. For example, assume that after fitting the log-linear model, the credible interval for an element of $\boldsymbol{\lambda}$ contains zero. When fitting the corresponding logistic regression model, the investigator will anticipate that the credible interval for the corresponding element of $\boldsymbol{\beta}$ will also contain zero. *Importantly*, for this translation to hold, it is essential that the prior distribution for $\boldsymbol{\beta}$ implied by the prior on $\boldsymbol{\lambda}$ is the same to the distribution the investigator would assign to $\boldsymbol{\beta}$ if they were to fit the logistic model directly. If the implied prior on $\boldsymbol{\beta}$ is not the same as a directly assigned prior then, with regard to $\boldsymbol{\beta}$, the correspondence from the Bayesian log-linear analysis to the logistic one becomes dubious. In both illustrations in Sect. 4, we observe that the credible intervals of the corresponding $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ parameters are virtually identical considering simulation error.

In Sect. 2, we provide the definition of the $g$-prior and mixtures of $g$-priors and describe how the $g$-prior is derived for log-linear and logistic regression models. Section 3 contains the main contributions in this manuscript. In Sect. 4, the correspondence from a log-linear to a logistic regression model is illustrated using simulated and real data. We conclude with a discussion.

## 2 The $g$-prior and mixtures of $g$-priors

A $g$-prior for the parameters $\boldsymbol{\gamma}$ of a generalized linear model is a multivariate Normal distribution $N(\boldsymbol{m}_\gamma, g\,\Sigma_\gamma)$, constructed so that the prior variance is a multiple of the inverse Fisher information matrix by a scalar $g$. See Liang et al. (2008) for a discussion on the choice of $g$. In accordance with Ntzoufras et al. (2003) and Ntzoufras (2009),

the $g$-prior for the parameters of log-linear and logistic regression models is specified so that, $\boldsymbol{m}_\gamma = (m_{\gamma_1}, 0, \ldots, 0)^\top$, where $m_{\gamma_1}$ corresponds to the intercept and can be nonzero, and,

$$\Sigma_\gamma = V(m^*)\zeta^{'}(m^*)^2 \left[ X_d^\top \text{diag}\left(\frac{1}{\phi_i}\right) X_d \right]^{-1},$$

where $\text{diag}(1/\phi_i)$ denotes a diagonal $n \times n$ matrix with nonzero elements $1/\phi_i$, and $m^* = \zeta^{-1}(m_{\gamma_1})$.

The unit information prior is a special case of the $g$-prior, obtained by setting $g = N$, where $N$ denotes the total number of observations. It is constructed so that the information contained in the prior is equal to the amount of information in a single observation (Kass and Wasserman 1995). Assuming that $g$ is a random variable, with prior $f(g)$, leads to a mixture of $g$-priors, so that,

$$\boldsymbol{\gamma}|g \sim N(\boldsymbol{m}_\gamma, g\Sigma_\gamma), \quad g \sim f(g).$$

Mixtures of $g$-priors are also called hyper-$g$ priors (Sabanès Bové and Held 2011).

*Log-linear regression* Consider counts $n_i$ $i = 1, \ldots, n_{ll}$. Now, $N = \sum_{i=1}^{n_{ll}} n_i$, and,

$$f(n_i|\mu_i) = \frac{\mathrm{e}^{-\mu_i}\mu_i^{n_i}}{n_i!},$$

with $\theta_i = \log(\mu_i)$, $b(\theta_i) = \mathrm{e}^{\theta_i}$ and $c(n_i, \phi_i) = -\log(n_i!)$. Also, $w_i\phi_i^{-1} = 1$, so that $w_i = 1$ implies $\phi_i = 1$. Note that,

$$\mu_i = b^{'}(\theta_i) = \mathrm{e}^{\theta_i}, \quad \text{Var}(n_i) = \phi_i w_i^{-1} b^{''}(\theta) = \mathrm{e}^{\theta_i}, \quad \text{and} \quad V(\mu_i) = \mu_i.$$

For the log-linear model, $\log(\boldsymbol{\mu}) = X_{ll}\boldsymbol{\lambda}$, and $\zeta(\mu_i) = \log(\mu_i)$ so that $\zeta^{'}(\mu_i) = \mu_i^{-1}$. The $g$-prior is constructed as $N(\boldsymbol{m}_\lambda, g\Sigma_\lambda)$, where $\boldsymbol{m}_\lambda = (\log(\bar{n}), 0, \ldots, 0)$. Here, $\bar{n}$ denotes the average cell count. The prior mean for the log-linear model intercept is also often set to zero (Dellaportas et al. 2012). (Note that altering the prior mean for the log-linear model intercept does not affect the validity of the theoretical results in Sect. 3. This is straightforward to deduce from the proof of Theorem 1 given in 'Appendix', as the prior mean for the log-linear intercept does not affect the implied distribution of the logistic regression parameters.) In addition,

$$\Sigma_\lambda = \bar{n}\frac{1}{(\bar{n})^2}\left(X_{ll}^\top X_{ll}\right)^{-1} = \frac{1}{\bar{n}}\left(X_{ll}^\top X_{ll}\right)^{-1} = \frac{n_{ll}}{N}\left(X_{ll}^\top X_{ll}\right)^{-1}.$$

*Logistic regression* Assume that $y_i$, $i = 1, \ldots, n_{lt}$, is the proportion of successes out of $t_i$ trials. Now, $N = \sum_{i=1}^{n_{lt}} t_i$, and,

$$f(t_i y_i|p_i) = \binom{t_i}{t_i y_i} p_i^{t_i y_i}(1-p_i)^{t_i - t_i y_i},$$

where $\theta_i = \text{logit}(p_i)$, $b(\theta_i) = \log(1 + e^{\theta_i})$, and $c(y_i, \phi_i) = \log\binom{t_i}{t_i y_i}$. Also, $w_i \phi_i^{-1} = t_i$, so that $w_i = 1$ implies $\phi_i = t_i^{-1}$. Note that,

$$E(y_i) = b^{'}(\theta_i) = \frac{e^{\theta_i}}{1 + e^{\theta_i}} = p_i, \quad \text{Var}(y_i) = \frac{\phi_i}{w_i} b^{''}(\theta_i) = \frac{1}{t_i} \frac{e^{\theta_i}}{(1 + e^{\theta_i})^2} = \frac{p_i(1 - p_i)}{t_i},$$

and,

$$V(p_i) = p_i(1 - p_i).$$

The logistic regression model is defined as $\text{logit}(\boldsymbol{p}) = X_{lt}\boldsymbol{\beta}$, so that $X_{lt}$ is a $n_{lt} \times n_\beta$ design matrix, and $\zeta(p_i) = \text{logit}(p_i)$ so that $\zeta^{'}(p_i) = [p_i(1 - p_i)]^{-1}$. The $g$-prior is $N(\boldsymbol{m}_\beta, g\Sigma_\beta)$, where $\boldsymbol{m}_\beta = (0, 0, \ldots, 0)$, and,

$$\Sigma_\beta = p^*(1 - p^*) \frac{1}{[p^*(1 - p^*)]^2} \left[ X_{lt}^\top \text{diag}(t_i) X_{lt} \right]^{-1} = \frac{1}{0.25} \left[ X_{lt}^\top \text{diag}(t_i) X_{lt} \right]^{-1}.$$

Here, $p^*$ corresponds to $m^*$ in the general definition of the $g$-prior at the start of this section, so that $p^* = \zeta^{-1}(m_{\gamma_1})$, where $m_{\gamma_1}$ is the first element of $\boldsymbol{m}_\beta$ which is zero. Thus, we obtain that $p^* = e^0/(e^0 + 1) = 0.5$. By approximating each $t_i$ with the average number of trials $\bar{t}$, as suggested by Ntzoufras et al. (2003),

$$\Sigma_\beta \simeq 4 \frac{1}{\bar{t}} \left( X_{lt}^\top X_{lt} \right)^{-1} = 4 \frac{n_{lt}}{\sum_{i=1}^{n_{lt}} t_i} \left( X_{lt}^\top X_{lt} \right)^{-1} = 4 \frac{n_{lt}}{N} \left( X_{lt}^\top X_{lt} \right)^{-1}.$$

## 3 Correspondence from log-linear to logistic regression models

Consider a set of categorical variables $\mathcal{P}$ that includes a binary variable $Y$. Assume a log-linear model that, in addition to the terms that involve $Y$, contains all possible interaction terms between the categorical factors in $\mathcal{P}\backslash\{Y\}$. We show that, given that a $g$-prior is assigned to the log-linear model parameters $\boldsymbol{\lambda}$, the implied prior for $\boldsymbol{\beta}$ is a $g$-prior for logistic regression models, i.e. the one that would be assigned if the investigator considered the logistic regression model directly.

**Theorem 1** *Assume a g-prior* $\boldsymbol{\lambda} \sim N(\boldsymbol{m}_\lambda, g\Sigma_\lambda)$ *on the parameters of a log-linear model* $log(\boldsymbol{\mu}) = X_{ll}\boldsymbol{\lambda}$, *that contains all possible interaction terms between the categorical factors in* $\mathcal{P}\backslash\{Y\}$. *This prior implies a g-prior* $N(\boldsymbol{m}_\beta, g\Sigma_\beta)$ *for the parameters* $\boldsymbol{\beta}$ *of the corresponding logistic regression* $logit(\boldsymbol{p}) = X_{lt}\boldsymbol{\beta}$.

*Proof* The proof is based on rearranging the rows and columns of $X_{ll}$, and partitioning so that one part of $X_{ll}$ consists of the logistic design matrix $X_{lt}$, or replications of $X_{lt}$. We then show that the prior mean and variance of the elements of $\boldsymbol{\lambda}$ that correspond to $\boldsymbol{\beta}$ are the prior that would be assigned to $\boldsymbol{\beta}$ if the logistic regression was fitted directly. The complete proof is given in 'Appendix'. □

**Corollary 1** *A unit information prior* $\boldsymbol{\lambda} \sim N(\boldsymbol{m}_\lambda, N\Sigma_\lambda)$ *implies a unit information prior* $N(\boldsymbol{m}_\beta, N\Sigma_\beta)$ *for the parameters* $\boldsymbol{\beta}$ *of the corresponding logistic regression.*

Corollary 1 follows directly from Theorem 1 by setting $g = N$. The following Corollary concerns mixtures of $g$-priors. It is implicitly assumed that the investigator would adopt the same prior density $f(g)$ for both modelling approaches.

**Corollary 2** *A mixture of g-priors so that* $\boldsymbol{\lambda}|g \sim N(\boldsymbol{m}_\lambda, g\Sigma_\lambda)$, $g \sim f(g)$, *implies a mixture of g-priors for the parameters* $\boldsymbol{\beta}$ *of the corresponding logistic regression, so that* $\boldsymbol{\beta}|g \sim N(\boldsymbol{m}_\beta, g\Sigma_\beta)$, $g \sim f(g)$.

This also follows from Theorem 1, which states that when $\boldsymbol{\lambda}|g \sim N(\boldsymbol{m}_\lambda, g\Sigma_\lambda)$, the conditional prior for $\boldsymbol{\beta}$ is $\boldsymbol{\beta}|g \sim N(\boldsymbol{m}_\beta, g\Sigma_\beta)$.

When the $g$-prior is utilized, it is common to assign a locally uniform Jeffreys prior ($\propto 1$) on the intercept, after the covariate columns of the design matrix have been centred to ensure orthogonality with the intercept (Liang et al. 2008). If one decides to adopt the approach where a flat prior is assigned to the intercept in both log-linear and logistic formulations, the correspondence between log-linear and logistic regression breaks, but only with regard to the intercept of the logistic regression. The prior on the log-linear intercept does not have a bearing on the implied prior for the logistic regression parameters, because the log-linear intercept does not contribute to the formation of the logistic regression parameters, as described in Sect. 1. After assigning a flat prior on the intercept of the log-linear model, all $\boldsymbol{\beta}$ parameters (including the intercept) are still Normal as linear combinations of Normal random variables, and the distribution of $\boldsymbol{\beta}$ is the one given by Theorem 1. For details, see the additional material in the proof of Theorem 1 in 'Appendix'. For an illustration, see Table 3 in Sect. 4.2.

Closed-form expressions for the posterior distribution of the parameters of a generalized linear model do not exist. However, it is known (O'Hagan and Forster 2004) that a Normal approximation applies. Consider a $g$-prior for the parameters $\boldsymbol{\gamma}$ of the generalized linear model, $\zeta(\boldsymbol{\mu}) = X_d\boldsymbol{\gamma}$, so that, for fixed $g$,

$$\boldsymbol{\gamma} \sim N(\boldsymbol{m}_\gamma, g\Sigma_\gamma).$$

Given observations $\boldsymbol{v} = \{v_1, \ldots, v_n\}$, the posterior distribution of $\gamma$ is approximated by a Normal density, so that,

$$\boldsymbol{\gamma}|\boldsymbol{v} \sim N\left(\left[g^{-1}\Sigma_\gamma^{-1} + \mathcal{I}(\hat{\boldsymbol{\gamma}})\right]^{-1} \times \left[g^{-1}\Sigma_\gamma^{-1}\boldsymbol{m}_\gamma + \mathcal{I}(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}\right], \left[g^{-1}\Sigma_\gamma^{-1} + \mathcal{I}(\hat{\boldsymbol{\gamma}})\right]^{-1}\right). \tag{1}$$

Here, $\hat{\boldsymbol{\gamma}}$ is the maximum likelihood estimate of $\boldsymbol{\gamma}$, and $\mathcal{I}(\hat{\boldsymbol{\gamma}})$ is the information matrix $X_d^\top \mathcal{V} X_d$. For the log-linear model, the diagonal matrix $\mathcal{V}$ (denoted by $\mathcal{V}_{\text{log-linear}}$) has diagonal elements $\exp\{X_{ll(i)}\hat{\boldsymbol{\lambda}}\}$, $i = 1, \ldots, n_{ll}$. When the logistic regression is fitted, $\mathcal{V}_{\text{logistic}}$ has diagonal elements $t_i \exp\{X_{lt(i)}\hat{\boldsymbol{\beta}}\}\exp\{1 + X_{lt(i)}\hat{\boldsymbol{\beta}}\}^{-2}$, $i = 1, \ldots, n_{lt}$. Within the Bayesian framework, when fitting a generalized linear model, a large sample ($n \to \infty$) will swamp the prior distribution, rendering it irrelevant for deriving

posterior inferences (O'Hagan and Forster 2004). In practice, this can be true even for moderate sample sizes (say, of order $10^2$ or larger), especially when the prior is not informative, which is typically the case with $g$-priors.

**Theorem 2** *Consider a g-prior* $\boldsymbol{\lambda} \sim N(\boldsymbol{m}_\lambda, g\Sigma_\lambda)$ *on the parameters of a log-linear model* $log(\boldsymbol{\mu}) = X_{ll}\boldsymbol{\lambda}$, *that contains all possible interaction terms between the categorical factors in* $\mathcal{P}\backslash\{Y\}$. *Consider also the analogous g-prior* $N(\boldsymbol{m}_\beta, g\Sigma_\beta)$ *for the parameters* $\boldsymbol{\beta}$ *of the corresponding logistic regression* $logit(\boldsymbol{p}) = X_{lt}\boldsymbol{\beta}$. *For fixed g, and for a large sample, the posterior distribution of* $\boldsymbol{\beta}$, *as given in* (1), *is approximately equal to the posterior distribution of the elements of* $\boldsymbol{\lambda}$ *that correspond to* $\boldsymbol{\beta}$.

*Proof* A partitioning similar to the one adopted for the proof of Theorem 1 is utilized. First, we show that, asymptotically, the posterior variance of $\boldsymbol{\beta}$ is the posterior variance of the elements of $\boldsymbol{\lambda}$ that correspond to $\boldsymbol{\beta}$. Then, we do the same for the posterior means. The proof is based on the crucial assumption that for a large sample the contribution of the prior in deriving the posterior moments can be ignored. A standard result utilized in the proof is that, asymptotically, the Binomial distribution for a data point can be approximated by a Poisson distribution. The complete proof is given in 'Appendix'. □

In the next section, we demonstrate with numerical illustrations that, for fixed $g$, the correspondence between the priors extends to posterior distributions, so that the posterior distribution of the logistic regression parameters matches the one of the corresponding log-linear model parameters. This is true even for relatively moderate sample sizes $N$, say a few hundred, and for standard choices of $g$ such as $g = N$.

## 4 Illustrations

Unit information priors were adopted for the model parameters ($g = N$). The size of the burn-in sample was $10^4$, followed by $5 \times 10^5$ iterations.

### 4.1 A simulation study

We simulate data from 1000 subjects, on six binary variables $\{Y, A, B, C, D, E\}$. Probabilities that correspond to the cells of the $2^6$ contingency table are generated in accordance with the log-linear model, $\log(\boldsymbol{\mu}) = YAB + YCD + YE$. Adopting the notation in Agresti (2002), a single letter denotes the presence of a main effect, two letter terms denote the presence of the implied first-order interaction and so on and so forth. The presence of an interaction between a set of variables implies the presence of all lower-order interactions plus main effects for that set. Cell counts are simulated according to the generated cell probabilities. Parameter values and the design matrix of the log-linear model used to generate the cell probabilities are given in Supplemental material, Section S2.

We fit to the simulated data the log-linear model,

$$\log(\boldsymbol{\mu}) = YAB + YCD + YE + ABCDE. \qquad \text{(M2)}$$

According to the discussion and results in Sects. 1 and 3, the corresponding logistic regression where $Y$ is treated as the outcome only contains the first-order interactions $AB$ and $CD$ plus the main effect for $E$,

$$\text{logit}(\boldsymbol{p}) = AB + CD + E. \tag{M3}$$

In Table 1, we present credible intervals (CIs) for the parameters of (M3) and the relevant parameters of (M2). The CIs for the corresponding $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ parameters are almost identical, considering simulation error. For example, the CI for $\lambda_{1,1,1}^{YCD}$ is $(-2.01, -0.85)$, whilst the CI for $\beta_{1,1}^{CD}$ is $(-2.00, -0.84)$.

In Table 2, we present minimum, maximum and quantile values for the $t_i$ observations, for the logistic regression in Table 1. It is clear that the simulated data do not represent balanced Binomial experiments where $t_i = \bar{t}$. The credible intervals listed in Table 1 demonstrate that the correspondence studied in this manuscript is very robust to departures from $t_i = \bar{t}$. This is also demonstrated in the real data analysis presented in the next subsection, where the collected data do not represent balanced Binomial experiments when one of the factors is treated as the outcome. In Supplemental material, we present additional analyses on simulated data sets, including results on smaller samples, roughly one quarter the size of the data set analysed in this section. Inferences on the correspondence between the posterior distributions remain unchanged.

### 4.2 A real data illustration

Edwards and Havránek (1985) presented a $2^6$ contingency table in which 1841 men were cross-classified by six binary risk factors $\{A, B, C, D, E, F\}$ for coronary heart disease. The data were also analysed in Dellaportas and Forster (1999), where the top hierarchical model was, $\log(\boldsymbol{\mu}) = AC + AD + AE + BC + CE + DE + F$, with posterior model probability 0.28. In Table 3, we present CIs for the parameters of the log-linear model,

$$AC + AD + AE + BCDEF. \tag{M4}$$

We also present CIs for the parameters of the corresponding logistic regression model when $A$ is treated as the outcome,

$$\text{logit}(\boldsymbol{p}) = C + D + E. \tag{M5}$$

We performed this analysis twice. Once after considering the $g$-priors described in Sect. 2 ($g = N$), as in the previous illustration, and after adopting a $g$-prior with a locally flat prior for the intercept. Under the $g$-prior described in Sect. 2, the CIs for the corresponding $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ parameters (including the intercept) are almost identical, considering simulation error. For instance, the CI for both the coefficient of $A$ in the log-linear model and the intercept in the logistic regression is $(-0.59, -0.24)$. Under the flat prior for the intercepts, the correspondence breaks down with regard to the intercept in the logistic regression model. The CI for the coefficient of $A$ in the log-linear model is $(-0.59, -0.24)$, whilst the CI for the intercept of the corresponding logistic regression model is $(-0.17, 0.02)$. Concurrently, the credible intervals for the

**Table 1** Simulated data illustration

| Log-linear model (M2), $\log(\boldsymbol{\mu}) = YAB + YCD + YE + ABCDE$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| Y | YA | YB | YC | YD | YE | YAB | YCD |
| (0.21, 1.07) | (−0.57,0.26) | (−0.44,0.43) | (−0.24,0.63) | (−0.38,0.50) | (−0.84,−0.27) | (−1.66,−0.50) | (−2.01,−0.85) |
| Outcome is Y (M3), $\text{logit}(\boldsymbol{p}) = AB + CD + E$ | | | | | | | |
| Intercept | A | B | C | D | E | AB | CD |
| (0.21, 1.08) | (−0.58, 0.27) | (−0.45, 0.43) | (−0.23, 0.61) | (−0.38, 0.49) | (−0.84, −0.27) | (−1.66, −0.50) | (−2.00, −0.84) |

Credible intervals (CIs) for the relevant parameters of log-linear model (M2), plus the parameters of the corresponding logistic regression (M3)

**Table 2** Simulated data illustration

| Outcome | Minimum | 25% quantile | Median | 75% quantile | Maximum |
|---|---|---|---|---|---|
| Y | 11 | 17 | 21 | 41.5 | 124 |
| A | 12 | 19 | 23 | 30 | 144 |
| B | 10 | 18 | 22.5 | 31 | 165 |
| C | 12 | 18.5 | 23 | 26.5 | 151 |
| D | 11 | 19.5 | 23 | 27.5 | 147 |
| E | 10 | 17.5 | 22 | 27 | 191 |

Maximum, minimum, and quantiles for $t_i, i = 1, \ldots, n_{lt}$, for different outcomes for the simulated data in Sect. 4.1

coefficients of $C$, $D$ and $E$ in the logistic regression model are almost identical to the corresponding CIs for $AC$, $AD$ and $AE$ in the log-linear model, with differences due to simulation error.

## 5 Discussion

The correspondence we investigated is not unexpected, given the results in Agresti (2002) discussed in Introduction, and also the link between the $g$-prior and Fisher's information matrix (Held et al. 2015), although this link is stronger for general linear models. Our investigation is also related to Consonni and Veronese (2008), where specifying a prior for the parameters of one model, and then, transferring this specification to the parameters of another is discussed. Of the four strategies considered in Consonni and Veronese (2008), the one directly linked to our manuscript is 'Marginalization', as the derived prior for the parameters of the logistic regression is the one that is the marginal prior of the relevant parameters of the log-linear model. Results on the relation between different statistical models are of interest, as they improve understanding and enhance the models' utility. Often, developments for one modelling framework are not readily available for the other. For example, Papathomas and Richardson (2016) comment on the relation between log-linear modelling and variable selection within clustering, in particular with regard to marginal independence, without examining logistic regression models.

Our numerical illustrations concern the $g$-prior, where the parameter $g$ is fixed. To further explore the correspondence between the two modelling frameworks, we also considered the two hyper priors that are prominent in Liang et al. (2008). This is the Zellner–Siow prior [IG(0.5, N/2)], and the prior introduced in the aforementioned manuscript in Sect. 4.2, with the suggested specification $\alpha = 3$. Furthermore, the two data sets were analysed after adopting a mixture of $g$-priors such that, $g \sim \text{IG}(a_g, b_g)$. We considered $a_g = 2 + \text{mean}(g)^2/\text{var}(g)$ and $b_g = \text{mean}(g) + \text{mean}(g)^3/\text{var}(g)$, in accordance with the specified prior moments mean$(g)$ and var$(g)$. We considered distinct Inverse Gamma densities with markedly different expectations and variances, as well as the vague prior $IG(0.1, 0.1)$. We observed that the correspondence does not hold exactly when a mixture of $g$-priors is adopted. This seems to be because the posterior distribution for $g$ is different under the two modelling frameworks, something that

**Table 3** Real data illustration

| Log-linear model (M4), $\log(\boldsymbol{\mu}) = AC + AD + AE + BCDEF$ ($g$-prior in Sect. 2) | | | |
|---|---|---|---|
| A | AC | AD | AE |
| $(-0.59, -0.24)$ | $(0.36, 0.74)$ | $(-0.56, -0.18)$ | $(0.30, 0.68)$ |
| Outcome is A (M5), $\text{logit}(\boldsymbol{p}) = C + D + E$ ($g$-prior in Sect. 2) | | | |
| Intercept | C | D | E |
| $(-0.59, -0.24)$ | $(0.37, 0.74)$ | $(-0.56, -0.18)$ | $(0.30, 0.68)$ |
| Log−linear model (M4), $\log(\boldsymbol{\mu}) = AC + AD + AE + BCDEF$ (flat prior on intercept) | | | |
| A | AC | AD | AE |
| $(-0.59, -0.24)$ | $(0.35, 0.76)$ | $(-0.55, -0.19)$ | $(0.29, 0.67)$ |
| Outcome is A (M5), $\text{logit}(\boldsymbol{p}) = C + D + E$ (flat prior on intercept) | | | |
| Intercept | C | D | E |
| $(-0.17, 0.02)$ | $(0.35, 0.75)$ | $(-0.56, -0.19)$ | $(0.30, 0.68)$ |

Relevant credible intervals for the parameters of log-linear model (M4) and the corresponding logistic regression model when A is treated as the outcome. Intervals are shown under the $g$-priors in Sect. 2 ($g = N$), and after considering a locally flat prior on the intercepts

affects to a small, but noticeable degree, the posterior credible intervals for the model parameters. For more details, see the analyses presented in Supplemental material.

Theoretical results in this manuscript refer to a specific log-linear model and the corresponding logistic regression model, for a given set of covariates. Therefore, our results should not be misinterpreted as licence to readily translate log-linear model selection inferences to inferences concerning logistic regression models. When performing model selection in a space of log-linear models, the prominent log-linear model describes a certain dependence structure between the categorical factors, including the relation of the binary Y with all other factors. The logistic regression that corresponds to the prominent log-linear model describes the dependence structure between $Y$ and the other factors that is supported by the data in accordance with the log-linear analysis. Therefore, under reasonable expectation, results from a single log-linear model determination analysis may translate, at the very least, to interesting logistic regressions for any of the binary factors that formed the contingency table. However, the mapping between log-linear and logistic regression model spaces is not bijective. Furthermore, posterior model probabilities depend on the prior on the model space, with various different approaches for defining such a prior discussed in Dellaportas et al. (2012). For the simulated data analysed in Sect. 4.1, log-linear model $YAB + YCD + YE$ has posterior probability 0.98, whilst the posterior probability of the corresponding logistic regression model (M3) is 0.59. Similar results from analysing the real data in Sect. 4.2, not presented here, also support this note of caution. In all model determination analyses, the Reversible Jump MCMC algorithm proposed in Papathomas et al. (2011) was employed. All possible graphical log-linear models were assumed equally likely a priori, as were all possible logistic graphical models for some given outcome.

**Acknowledgements** The author wishes to thank Professor Petros Dellaportas and Dr. Antony Overstall for useful discussions during the preparation of this manuscript. We would also like to thank two reviewers and the editors for comments that helped to improve the manuscript.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## Appendix

*Proof of Theorem 1* To facilitate the proof, the following notation is introduced. Using the incidence matrix $\boldsymbol{T}$ discussed in Sect. 1, write the mapping between $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ as $\boldsymbol{\beta} = \boldsymbol{T}\boldsymbol{\lambda}$, where

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{\lambda}_{(1)} \\ \vdots \\ \boldsymbol{\lambda}_{(n_{\lambda_Y})} \end{pmatrix},$$

and $\boldsymbol{\lambda}_{(k)}, k = 1, \ldots, n_{\lambda_Y}$, is a vector of zeros with the exception of one element that is equal to one. This element is in the position of the $k$th $\boldsymbol{\lambda}$ parameter with a $Y$ in its superscript. With $n_{\lambda_Y}$ we denote the number of parameters in $\boldsymbol{\lambda}$ with a $Y$ in their

superscript. This is a more rigorous definition of $\boldsymbol{T}$ compared to the more descriptive definition in Sect. 1. To ease algebraic calculations, and without any loss of generality, rearrange the columns of $\boldsymbol{\lambda}$, creating a new vector $\boldsymbol{\lambda}_r$, so that $\boldsymbol{T}$ changes accordingly to, $\boldsymbol{T}_r = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \end{pmatrix}$, where $\boldsymbol{I}$ is an $n_\beta \times n_\beta$ identity matrix and $n_\beta$ is the number of elements in $\boldsymbol{\beta}$. The rows and columns of $X_{ll}$ are also rearranged accordingly to create $X_{rll}$, so that,

$$X_{rll} = \begin{pmatrix} X^*_{lt} & X_{ll\text{-}lt} \\ \boldsymbol{0} & X_{ll\text{-}lt} \end{pmatrix} \tag{2}$$

$X_{ll\text{-}lt}$ is a square $(n_{ll}/2 \times n_{ll}/2)$ matrix. This is because we consider the log-linear model that, in addition to the terms that involve $Y$, contains all possible interaction terms between the categorical factors in $\mathcal{P}\backslash\{Y\}$. The number of parameters that correspond to the intercept, main effects and interactions for $\mathcal{P}\backslash\{Y\}$ is $n_{ll}/2$.

Denote with $j_1 = 2$ the number of levels of the binary factor $Y$ that becomes the outcome in the logistic regression model. With $j_2$ to $j_q$, $1 \leq q \leq P-1$ denote the number of levels of the $q-1$ factors that are present in the log-linear model but disappear from the logistic regression model as they do not interact with $Y$. Then, $n_{ll} = 2 \times j_2 \times \cdots \times j_q \times n_{lt}$. When $q = 1$, all factors other than $Y$ remain in the logistic regression model as covariates. When $q = P-1$, the corresponding logistic regression model only contains the intercept. For instance, for a $2^P$ contingency table, $n_{ll} = 2^q \times n_{lt}$, and for $q = 1, n_{ll} = 2 \times n_{lt}$. Furthermore, $X^*_{lt}$ is a $n_{ll}/2 \times n_\beta$ matrix. By rearranging the rows of $X_{rll}$ when necessary, we can write $X^*_{lt}$ as $X^*_{lt} = (X_{lt}^\top X_{lt}^\top \ldots X_{lt}^\top)^\top$, where $X_{lt}^\top$ is repeated $(j_1 - 1) \times j_2 \times \cdots \times j_q$ times. For example, for $q = 1$, $X^*_{lt} = X_{lt}$. For $q = 2$, $X_{lt}$ repeats $j_2$ times within $X^*_{lt}$.

We can now write $\boldsymbol{\beta} = \boldsymbol{T}_r \boldsymbol{\lambda}_r$. For example, assume the log-linear model (M1) describes a $3 \times 2 \times 2$ contingency table. Then, $q = 1$, and the standard arrangement of the elements of $\boldsymbol{\lambda}$ would be such that,

$$X_{ll} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{\lambda} = \begin{pmatrix} \lambda \\ \lambda_1^X \\ \lambda_2^X \\ \lambda_1^Y \\ \lambda_1^Z \\ \lambda_{11}^{XY} \\ \lambda_{21}^{XY} \\ \lambda_{11}^{XZ} \\ \lambda_{21}^{XZ} \\ \lambda_{11}^{YZ} \end{pmatrix},$$

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

After rearranging,

$$X_{rll} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\lambda}_r = \begin{pmatrix} \lambda_1^Y \\ \lambda_{11}^{XY} \\ \lambda_{21}^{XY} \\ \lambda_{11}^{YZ} \\ \lambda \\ \lambda_1^X \\ \lambda_2^X \\ \lambda_1^Z \\ \lambda_{11}^{XZ} \\ \lambda_{21}^{XZ} \end{pmatrix},$$

$$\boldsymbol{T}_r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

For another example, where $q = 2$, consider again model (M1) but now assume that the interaction $YZ$ is not present in the log-linear model. Then, the $Z$ factor will disappear from the corresponding logistic regression model, and after rearranging,

$$X_{rll} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\lambda}_r = \begin{pmatrix} \lambda_1^Y \\ \lambda_{11}^{XY} \\ \lambda_{21}^{XY} \\ \lambda \\ \lambda_1^X \\ \lambda_2^X \\ \lambda_1^Z \\ \lambda_{11}^{XZ} \\ \lambda_{21}^{XZ} \end{pmatrix},$$

$$\boldsymbol{T}_r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The $g$-prior,

$$\boldsymbol{\lambda} \sim N(\boldsymbol{m}_\lambda, g\,\Sigma_\lambda) \equiv N\left((\log(\bar{n}), 0, \ldots, 0)^\top, \frac{g n_{ll}}{N}\left(X_{ll}^\top X_{ll}\right)^{-1}\right),$$

translates to,

$$\boldsymbol{\lambda}_r \sim N(\boldsymbol{m}_{\lambda_r}, g\Sigma_{\lambda_r}) \equiv N\left((0,\ldots,0,\log(\bar{n}),0,\ldots,0)^\top, \frac{gn_{ll}}{N}\left(X_{rll}^\top X_{rll}\right)^{-1}\right),$$

where $\log(\bar{n})$ is the $(n_\beta+1)$th element in the mean vector. Then,

$$E(\boldsymbol{\beta}) = E(\boldsymbol{T}_r\boldsymbol{\lambda}_r) = \boldsymbol{T}_r E(\boldsymbol{\lambda}_r) = \begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\end{pmatrix} \times \boldsymbol{\mu}_{\lambda_r} = \boldsymbol{0}.$$

Furthermore,

$$\text{Var}(\boldsymbol{\beta}) = g\boldsymbol{T}_r\Sigma_{\lambda_r}\boldsymbol{T}_r^\top = \frac{gn_{ll}}{N}\boldsymbol{T}_r\left(X_{rll}^\top X_{rll}\right)^{-1}\boldsymbol{T}_r^\top.$$

From (2),

$$\begin{aligned}\left(X_{rll}^\top X_{rll}\right)^{-1} &= \begin{pmatrix} X_{lt}^{*\top}X_{lt}^* & X_{lt}^{*\top}X_{ll\text{-}lt} \\ X_{ll\text{-}lt}^\top X_{lt}^* & X_{ll\text{-}lt}^\top X_{ll\text{-}lt} + X_{ll\text{-}lt}^\top X_{ll\text{-}lt}\end{pmatrix}^{-1} \\ &= \begin{pmatrix} X_{lt}^{*\top}X_{lt}^* & X_{lt}^{*\top}X_{ll\text{-}lt} \\ X_{ll\text{-}lt}^\top X_{lt}^* & 2X_{ll\text{-}lt}^\top X_{ll\text{-}lt}\end{pmatrix}^{-1}.\end{aligned}$$

From Lutkepohl (1996, p. 147), the submatrix $H$ that is formed by the first $n_\beta$ rows and columns of $(X_{rll}^\top X_{rll})^{-1}$ is,

$$\begin{aligned}H = &\left(X_{lt}^{*\top}X_{lt}^*\right)^{-1} + \left(X_{lt}^{*\top}X_{lt}^*\right)^{-1}X_{lt}^{*\top}X_{ll\text{-}lt}\left[X_{ll\text{-}lt}^\top\left(2\boldsymbol{I} - X_{lt}^*\left(X_{lt}^{*\top}X_{lt}^*\right)^{-1}X_{lt}^{*\top}\right)X_{ll\text{-}lt}\right]^{-1} \\ &\times X_{ll\text{-}lt}^\top X_{lt}^*\left(X_{lt}^{*\top}X_{lt}^*\right)^{-1}.\end{aligned}$$

Now, $P_{lt} \equiv X_{lt}^*(X_{lt}^{*\top}X_{lt}^*)^{-1}X_{lt}^{*\top}$ is the projection matrix for $X_{lt}^*$. It is straightforward to verify that for a projection matrix $P_{lt}$ and a constant $c$,

$$(c\boldsymbol{I} - P_{lt}) \times \left(\frac{1}{c}\boldsymbol{I} + \frac{1}{c(c-1)}P_{lt}\right) = \boldsymbol{I}.$$

Therefore, $(2\boldsymbol{I} - P_{lt}) = (0.5\boldsymbol{I} + 0.5P_{lt})^{-1}$, and consequently,

$$\begin{aligned}H = &\left(X_{lt}^{*\top}X_{lt}^*\right)^{-1} + \left(X_{lt}^{*\top}X_{lt}^*\right)^{-1}X_{lt}^{*\top}X_{ll\text{-}lt}\left[X_{ll\text{-}lt}^\top(0.5\boldsymbol{I} + 0.5P_{lt})^{-1}X_{ll\text{-}lt}\right]^{-1} \\ &\times X_{ll\text{-}lt}^\top X_{lt}^*\left(X_{lt}^{*\top}X_{lt}^*\right)^{-1}.\end{aligned}$$

$X_{ll\text{-}lt}$ is a square matrix of full rank. If $X_{ll\text{-}lt}$ was not full rank, then some of its columns would be linearly dependent. In turn, some of the columns of $\begin{pmatrix} X_{ll\text{-}lt} \\ X_{ll\text{-}lt}\end{pmatrix}$ would

be linearly dependent, implying the same for columns of $X_{rll}$ (see Eq. 2). This is not possible as $X_{rll}$ is a design matrix of full rank. Thus, $X_{ll\text{-}lt}^{-1}$ exists and,

$$\begin{aligned}
H &= \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} + \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} X_{lt}^{*\top} X_{ll\text{-}lt} \left[X_{ll\text{-}lt}^{-1} \left(0.5\boldsymbol{I} + 0.5P_{lt}\right) X_{ll\text{-}lt}^{\top^{-1}}\right] \\
&\quad \times X_{ll\text{-}lt}^{\top} X_{lt}^{*} \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} \\
&= \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} + \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} X_{lt}^{*\top} \left(0.5\boldsymbol{I} + 0.5P_{lt}\right) X_{lt}^{*} \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} \\
&= \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} + 0.5 \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} + 0.5 \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} \\
&= 2 \left(X_{lt}^{*\top} X_{lt}^{*}\right)^{-1} \\
&= 2 \left(j_2 \times \cdots \times j_q X_{lt}^{\top} X_{lt}\right)^{-1}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{Var}(\boldsymbol{\beta}) &= \frac{gn_{ll}}{N} \boldsymbol{T}_r \left(X_{rll}^{\top} X_{rll}\right)^{-1} \boldsymbol{T}_r^{\top} \\
&= \frac{gn_{ll}}{N} \left(\boldsymbol{I}\ \boldsymbol{0}\right) \left(X_{rll}^{\top} X_{rll}\right)^{-1} \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} \\
&= \frac{2g2j_2 \times \cdots \times j_q n_{lt}}{Nj_2 \times \cdots \times j_q} \left(X_{lt}^{\top} X_{lt}\right)^{-1} \\
&= \frac{4gn_{lt}}{N} \left(X_{lt}^{\top} X_{lt}\right)^{-1}
\end{aligned}$$

Thus,

$$\boldsymbol{\beta} \sim N\left(\boldsymbol{0}, \frac{4gn_{lt}}{N} \left(X_{lt}^{\top} X_{lt}\right)^{-1}\right),$$

which is the $g$-prior for the parameters of a logistic regression, as described in Sect. 2. This completes the proof. □

*Placing a flat prior on the intercept* Assume that a flat prior is placed on the intercept of the log-linear model, after the design matrix has been centred to induce orthogonality between the intercept and the factors that form the contingency table. This does not alter the prior on the parameters of the corresponding logistic regression model. The proof follows along the lines of the proof of Theorem 1, if we express the parameters of the logistic regression model as $\boldsymbol{\beta} = \boldsymbol{T}_{r-1}\boldsymbol{\lambda}_{r-1}$, where $\boldsymbol{T}_{r-1}$ denotes matrix $\boldsymbol{T}_r$ without the first column with all elements zero, and $\boldsymbol{\lambda}_{r-1}$ denotes the vector of parameters $\boldsymbol{\lambda}_r$ without the intercept $\lambda$. The proof proceeds as above, replacing $X_{rll}$ with $X_{rll-1}$, where $X_{rll-1}$ is the former matrix without the column with all elements one. It is also required to replace $X_{ll\text{-}lt}$ with $X_{ll\text{-}lt-1}$, where $X_{ll\text{-}lt-1}$ is the former matrix without the column with all elements one.

*Proof of Theorem 2* The proof utilizes quantities defined earlier in Sect. 3 and in the proof of Theorem 1. First, we will show that, asymptotically, the posterior variance of $\boldsymbol{\beta}$ is identical to the posterior variance of the elements of $\boldsymbol{\lambda}$ that correspond to $\boldsymbol{\beta}$. Then, we will do the same for the posterior means.

Consider a vector of cell counts $\boldsymbol{n} = \{n_1, \ldots, n_{ll}\}$, and the log-linear model $\log(\boldsymbol{\mu}) = X_{ll}\boldsymbol{\lambda}$. Then, asymptotically,

$$\begin{aligned}\text{Var}(\boldsymbol{\lambda}|\boldsymbol{n}) &\simeq \left[g^{-1}\Sigma_{\lambda}^{-1} + \mathcal{I}(\hat{\boldsymbol{\lambda}})\right]^{-1} \\ &= \left[\frac{N}{gn_{ll}} X_{ll}^{\top} X_{ll} + X_{ll}^{\top}\mathcal{V}(\hat{\boldsymbol{\lambda}})X_{ll}\right]^{-1},\end{aligned}$$

where $\hat{\boldsymbol{\lambda}}$ denotes the maximum likelihood estimate (MLE). After rearranging the rows and columns of $X_{ll}$, consider the log-linear model with linear predictor $X_{rll}\boldsymbol{\lambda}_r$, for cell counts $\boldsymbol{n}_r$, where $\boldsymbol{n}_r$ is $\boldsymbol{n}$ rearranged to correspond to $X_{rll}$. Now,

$$\begin{aligned}\text{Var}(\boldsymbol{\lambda}_r|\boldsymbol{n}_r) &\simeq \left[g^{-1}\Sigma_{\lambda_r}^{-1} + \mathcal{I}\left(\hat{\boldsymbol{\lambda}}_r\right)\right]^{-1} \\ &= \left[\frac{N}{gn_{ll}} X_{rll}^{\top} X_{rll} + X_{rll}^{\top}\mathcal{V}(\hat{\boldsymbol{\lambda}}_r)X_{rll}\right]^{-1} \\ &= \left[X_{rll}^{\top}\left(\frac{N}{gn_{ll}} + \mathcal{V}(\hat{\boldsymbol{\lambda}_r})\right)X_{rll}\right]^{-1} \\ &= \left[\left(\begin{pmatrix} \frac{N}{gn_{ll}}\boldsymbol{I} + \mathcal{V}_1\mathcal{V}_2 & \mathbf{0} \\ \mathbf{0} & \frac{N}{gn_{ll}} + \mathcal{V}_2 \end{pmatrix}^{1/2} \begin{pmatrix} X_{lt}^{*} & X_{ll\text{-}lt} \\ \mathbf{0} & X_{ll\text{-}lt} \end{pmatrix}\right)^{\top}\right. \\ &\quad \left.\times \begin{pmatrix} \frac{N}{gn_{ll}}\boldsymbol{I} + \mathcal{V}_1\mathcal{V}_2 & \mathbf{0} \\ \mathbf{0} & \frac{N}{gn_{ll}} + \mathcal{V}_2 \end{pmatrix}^{1/2} \begin{pmatrix} X_{lt}^{*} & X_{ll\text{-}lt} \\ \mathbf{0} & X_{ll\text{-}lt} \end{pmatrix}\right]^{-1}.\end{aligned}$$

$\mathcal{V}_1$ denotes a diagonal matrix with nonzero elements $\exp(X_{lt(i)}^{*}(\boldsymbol{T}_r\hat{\boldsymbol{\lambda}}_r))$, $i = 1, \ldots, n_{ll}/2$. $\mathcal{V}_2$ denotes a diagonal matrix with nonzero elements $\exp(X_{ll\text{-}lt(i)}\hat{\boldsymbol{\lambda}}_{ll\text{-}lt})$, $i = 1, \ldots, n_{ll}/2$, where $\hat{\boldsymbol{\lambda}}_{ll\text{-}lt}$ denotes the MLE for $\boldsymbol{\lambda}_r \backslash \boldsymbol{T}_r\boldsymbol{\lambda}_r$. Now,

$$\text{Var}(\boldsymbol{\lambda}_r|\boldsymbol{n}_r) \simeq \begin{pmatrix} X_{lt}^{*\top} A_{12} X_{lt}^{*} & X_{lt}^{*\top} A_{12} X_{ll\text{-}lt} \\ X_{ll\text{-}lt}^{\top} A_{12} X_{lt}^{*} & X_{ll\text{-}lt}^{\top}(A_{12} + A_2)X_{ll\text{-}lt} \end{pmatrix}^{-1},$$

where $A_{12} = \frac{N}{gn_{ll}}\boldsymbol{I} + \mathcal{V}_1\mathcal{V}_2$ and $A_2 = \frac{N}{gn_{ll}}\boldsymbol{I} + \mathcal{V}_2$. From Lutkepohl (1996, p. 147), the submatrix $H$ that is formed by the first $n_\beta$ rows and columns of $\text{Var}(\boldsymbol{\lambda}_r|\boldsymbol{n}_r)$ is,

$$H = \left(X_{lt}^{*\top} A_{12} X_{lt}^{*}\right)^{-1} + \left(X_{lt}^{*\top} A_{12} X_{lt}^{*}\right)^{-1} X_{lt}^{*\top} A_{12} X_{ll\text{-}lt}$$

$$
\begin{aligned}
&\times \left[ X_{ll\text{-}lt}^{\top}(A_{12}+A_2)X_{ll\text{-}lt} - X_{ll\text{-}lt}^{\top}A_{12}X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}X_{ll\text{-}lt}\right]^{-1} \\
&\times X_{ll\text{-}lt}^{\top}A_{12}X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1} \\
&= \left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1} \\
&+ \left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}\left[(A_{12}+A_2) - A_{12}X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}\right]^{-1} \\
&\times A_{12}X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1} \\
&= \left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1} \\
&+ \left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}\left[(\boldsymbol{I}+A_{12}^{-1}A_2) - X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}\right]^{-1} \\
&\times X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1} \\
&= \left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1} \\
&+ \left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}\left[\boldsymbol{I} - \left(\boldsymbol{I}+A_{12}^{-1}A_2\right)^{-1}X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}A_{12}\right]^{-1} \\
&\times X_{lt}^{*}\left(X_{lt}^{*\top}A_{12}X_{lt}^{*}\right)^{-1}.
\end{aligned}
$$

From Lutkepohl (1996, p. 29, line 6), the expression above simplifies to,

$$
\begin{aligned}
H &= \left(X_{lt}^{*\top}A_{12}X_{lt}^{*} - X_{lt}^{*\top}A_{12}\left(\boldsymbol{I}+A_{12}^{-1}A_2\right)^{-1}X_{lt}^{*}\right)^{-1} \\
&= \left[X_{lt}^{*\top}\left(A_{12} - A_{12}(\boldsymbol{I}+A_{12}^{-1}A_2)^{-1}\right)X_{lt}^{*}\right]^{-1}.
\end{aligned}
$$

Within the Bayesian framework, a large sample ($N \to \infty$) will swamp the prior distribution, rendering it irrelevant for deriving posterior inferences (O'Hagan and Forster 2004). This can be viewed as equivalent to considering a flat non-informative prior, in our case assuming that $g \to \infty$. For a sample size large enough to justify ignoring the contribution of the prior distribution in $\mathrm{Var}(\boldsymbol{\lambda}|\boldsymbol{n})$, i.e. assuming that $A_{12} = \mathcal{V}_1\mathcal{V}_2$ and $A_2 = \mathcal{V}_2$, asymptotically,

$$
\begin{aligned}
H &\simeq \left[X_{lt}^{*\top}\left(\mathcal{V}_1\mathcal{V}_2 - \mathcal{V}_1\mathcal{V}_2(\boldsymbol{I}+\mathcal{V}_1^{-1}\mathcal{V}_2^{-1}\mathcal{V}_2)^{-1}\right)X_{lt}^{*}\right]^{-1} \\
&= \left[X_{lt}^{*\top}\left(\mathcal{V}_1\mathcal{V}_2 - \mathcal{V}_1^{2}\mathcal{V}_2(\boldsymbol{I}+\mathcal{V}_1)^{-1}\right)X_{lt}^{*}\right]^{-1} \\
&= \left[X_{lt}^{*\top}\left[\left(\mathcal{V}_1\mathcal{V}_2(\boldsymbol{I}+\mathcal{V}_1) - \mathcal{V}_1^{2}\mathcal{V}_2\right)(\boldsymbol{I}+\mathcal{V}_1)^{-1}\right]X_{lt}^{*}\right]^{-1} \\
&= \left[X_{lt}^{*\top}\left(\mathcal{V}_1\mathcal{V}_2(\boldsymbol{I}+\mathcal{V}_1)^{-1}\right)X_{lt}^{*}\right]^{-1} \\
&= \left[X_{lt}^{\top}(\mathcal{V}_{1,\text{reduced}}(\boldsymbol{I}+\mathcal{V}_{1,\text{reduced}})^{-1}\left[\mathcal{V}_{2,1}+\mathcal{V}_{2,2}+\cdots+\mathcal{V}_{2,(j_1-1)\times j_2\times\cdots\times j_q}\right]X_{lt}\right]^{-1}.
\end{aligned}
$$

$\mathcal{V}_{1,\text{reduced}}$ denotes a diagonal matrix with elements $\exp(X_{lt(i)}(\boldsymbol{T}_r\hat{\boldsymbol{\lambda}}_r))$, $i = 1,\ldots,n_{lt}$. $\mathcal{V}_{2,k}$, $k = 1,\ldots,(j_1-1)\times j_2\times\cdots\times j_q$, denotes a diagonal matrix with elements $\exp(X_{ll\text{-}lt(n_{lt}(k-1)+i)}\hat{\boldsymbol{\lambda}}_{ll\text{-}lt})$. This expression simplifies as $q$ becomes smaller, i.e. the fewer times $X_{lt}$ is contained within $X^*_{lt}$. For example, when $X^*_{lt} = X_{lt}$, i.e. when $q = 1$ and all factors other than $Y$ remain in the logistic regression, $\mathcal{V}_{1,\text{reduced}} = \mathcal{V}_1$.

We now utilize the standard result (see, for example, Rohatgi 1976, p. 200) that, asymptotically, the Binomial distribution $\text{Bin}(t_i, \frac{\exp(X^*_{lt(i)}(\boldsymbol{T}_r\boldsymbol{\lambda}_r))}{1+\exp(X^*_{lt(i)}(\boldsymbol{T}_r\boldsymbol{\lambda}_r))})$ of a data point $t_i\,y_i$, $i = 1,\ldots,n_{lt}$, can be approximated by a Poisson distribution $\text{Poisson}(t_i\frac{\exp(X^*_{lt(i)}(\boldsymbol{T}_r\boldsymbol{\lambda}_r))}{1+\exp(X^*_{lt(i)}(\boldsymbol{T}_r\boldsymbol{\lambda}_r))})$. The Binomial observation $t_i - t_i\times y_i$ is formed by adding $(j_1-1)\times j_2\times\cdots\times j_q$ independent Poisson cell counts. Considering the Poisson log-linear model, $t_i - t_i\,y_i$ follows the Poisson distribution,

$$\text{Poisson}(\exp(X_{ll\text{-}lt(i)}\hat{\boldsymbol{\lambda}}_{ll\text{-}lt}) + \cdots + \exp(X_{ll\text{-}lt(n_{lt}((j_1-1)\times j_2\times\cdots\times j_q-1)+i)}\hat{\boldsymbol{\lambda}}_{ll\text{-}lt})).$$

Therefore, approximately,

$$\begin{aligned} & t_i\frac{1}{1+\exp(X_{lt(i)}(\boldsymbol{T}_r\hat{\boldsymbol{\lambda}}_r))} \\ & \simeq \exp(X_{ll\text{-}lt(i)}\hat{\boldsymbol{\lambda}}_{ll\text{-}lt}) + \cdots + \exp(X_{ll\text{-}lt(n_{lt}((j_1-1)\times j_2\times\cdots\times j_q-1)+i)}\hat{\boldsymbol{\lambda}}_{ll\text{-}lt}). \end{aligned} \quad (3)$$

In matrix notation, we can now write that, asymptotically,

$$\begin{aligned} \text{Var}(\boldsymbol{T}_r\boldsymbol{\lambda}_r|\boldsymbol{n}_r) &= \boldsymbol{T}_r(\text{Var}(\boldsymbol{\lambda}_r|\boldsymbol{n}_r))\boldsymbol{T}_r^\top \\ &= \begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\end{pmatrix}(\text{Var}(\boldsymbol{\lambda}_r|\boldsymbol{n}_r))\begin{pmatrix}\boldsymbol{I}\\ \boldsymbol{0}\end{pmatrix} \\ &\simeq \left[X_{lt}^\top\left(\boldsymbol{t}\mathcal{V}_{1,\text{reduced}}(\boldsymbol{I}+\mathcal{V}_{1,\text{reduced}})^{-2}\right)X_{lt}\right]^{-1} \\ &= \left(X_{lt}^\top\mathcal{V}_{\text{logistic}}X_{lt}\right)^{-1}, \end{aligned}$$

where $\boldsymbol{t}$ is a diagonal matrix with diagonal elements the number of trials $t_i$, and $\mathcal{V}_{\text{logistic}}$ has diagonal elements $t_i\exp\{X_{lt(i)}\hat{\boldsymbol{\beta}}\}\exp\{1 + X_{lt(i)}\hat{\boldsymbol{\beta}}\}^{-2}$, $i = 1,\ldots,n_{lt}$. $(X_{lt}^\top\mathcal{V}_{\text{logistic}}X_{lt})^{-1}$ is, asymptotically, the posterior variance of $\boldsymbol{\beta}$ when the logistic regression is fitted directly, and thus, we have shown that the posterior variance of $\boldsymbol{\beta}$ is identical to the posterior variance of the elements of $\boldsymbol{\lambda}$ that correspond to $\boldsymbol{\beta}$.

We will now show that, asymptotically, the posterior mean $E(\boldsymbol{\beta}|\boldsymbol{t},\boldsymbol{y})$ is the posterior mean of the elements of $\boldsymbol{\lambda}$ that correspond to $\boldsymbol{\beta}$. For a sample large enough to justify ignoring the contribution of the prior in (1), we obtain that, $E(\boldsymbol{\lambda}|\boldsymbol{n}) \simeq \mathcal{I}(\hat{\boldsymbol{\lambda}})^{-1}\mathcal{I}(\hat{\boldsymbol{\lambda}})\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}$. Similarly, $E(\boldsymbol{\beta}|\boldsymbol{t},\boldsymbol{y}) \simeq \hat{\boldsymbol{\beta}}$. Therefore, $E(\boldsymbol{T}_r\boldsymbol{\lambda}_r|\boldsymbol{n}) \simeq \boldsymbol{T}_r\hat{\boldsymbol{\lambda}}_r$, and it is sufficient to show that $\hat{\boldsymbol{\beta}} = \boldsymbol{T}_r\hat{\boldsymbol{\lambda}}_r$. Closed-form expressions for the maximum likelihood estimators of the parameters of a generalized linear model do not exist. As a result, we will base the derivation of this result on the Iterative Re-weighed Least Squares (IRLS) algorithm. This is the standard procedure for maximizing the likelihood when a generalized linear

model is fitted. See Wood (2006) for more details. For a linear predictor $X_d\boldsymbol{\gamma}$, this iterative process is based on the formula,

$$\boldsymbol{\gamma}^{it+1} = \boldsymbol{\gamma}^{it} + \left(X_d^\top \mathcal{V}(\boldsymbol{\gamma}^{it}) X_d\right)^{-1} X_d^\top \mathcal{V}(\boldsymbol{\gamma}^{it}) \boldsymbol{\zeta}^{it}.$$

For a log-linear model, $\boldsymbol{\zeta}^{it}$ is denoted by $\boldsymbol{\zeta}^{it}_{\text{log-linear}}$, and its $i$th element, $i = 1, \ldots, n_{ll}$, is,

$$\zeta_{\text{log-linear}(i)} = \frac{n_i}{\exp(X_{rll(i)}\boldsymbol{\lambda}_r^{it})} - 1.$$

For a logistic regression model, $\boldsymbol{\zeta}^{it}$ is denoted by $\boldsymbol{\zeta}^{it}_{\text{logistic}}$, and its $i$th element, $i = 1, \ldots, n_{lt}$, is,

$$\zeta_{\text{logistic}(i)} = \frac{t_i y_i \left(1 + \exp(X_{lt}\boldsymbol{\beta}^{it})\right) - t_i \exp(X_{lt}\boldsymbol{\beta}^{it})}{t_i} \frac{1 + \exp(X_{lt}\boldsymbol{\beta}^{it})}{\exp(X_{lt}\boldsymbol{\beta}^{it})}.$$

For the log-linear model, the IRLS procedure is written as,

$$\boldsymbol{\lambda}_r^{it+1} = \boldsymbol{\lambda}_r^{it} + \left(X_{rll}^\top \mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}_r^{it}) X_{rll}\right)^{-1} X_{rll}^\top \mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}_r^{it}) \boldsymbol{\zeta}^{it}_{\text{log-linear}},$$

where $\mathcal{V}_{\text{log-linear}}$ is a diagonal matrix with diagonal elements $\exp\{X_{rll(i)}\hat{\boldsymbol{\lambda}}_r\}, i = 1, \ldots, n_{ll}$. Algebraic operations similar to the ones carried out earlier show that $(X_{rll}^\top \mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}^{it}) X_{rll})^{-1}$ partitions as,

$$\begin{pmatrix} \left(X_{lt}^\top \mathcal{V}_{\text{logistic}} X_{lt}\right)^{-1} & -\left[X_{lt}^{*\top} \mathcal{V}_1 \mathcal{V}_2 X_{lt}^*\right]^{-1} X_{lt}^{*\top} \mathcal{V}_1 \times \left[\mathcal{V}_1 + \boldsymbol{I} - \mathcal{V}_1 \mathcal{V}_2 X_{lt}^*\right. \\ & \left. \times \left[X_{lt}^{*\top} \mathcal{V}_1 \mathcal{V}_2 X_{lt}^*\right]^{-1} X_{lt}^{*\top} \mathcal{V}_1\right]^{-1} X_{ll\text{-}lt}^{\top -1} \\ \Omega_1 & \Omega_2 \end{pmatrix},$$

where $\Omega_1$ and $\Omega_2$ are matrices not relevant to this proof. Furthermore, $X_{rll}^\top \mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}_r^{it})$ partitions as,

$$\begin{pmatrix} X_{lt}^{*\top} \mathcal{V}_1 \mathcal{V}_2 & \mathbf{0} \\ X_{ll\text{-}lt}^\top \mathcal{V}_1 \mathcal{V}_2 & X_{ll\text{-}lt}^\top \mathcal{V}_2 \end{pmatrix}.$$

For the log-linear model, we write $\boldsymbol{\zeta}_{\text{log-linear}} = (\boldsymbol{\zeta}_{lt}^{*\top} \boldsymbol{\zeta}_{ll\text{-}lt}^\top)^\top$, where $\boldsymbol{\zeta}_{lt}^*$ corresponds to the first $n_{ll}/2$ rows of $X_{rll}$. Now, the first $n_\beta$ elements of $(X_{rll}^\top \mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}^{it}) X_{rll})^{-1} X_{rll}^\top \mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}_r^{it}) \boldsymbol{\zeta}_{\text{log-linear}}$, i.e. the ones that correspond to the logistic regression parameters, are given by,

$$\left(X_{lt}^{\top}\mathcal{V}_{\text{logistic}}X_{lt}\right)^{-1}X_{lt}^{*\top}\mathcal{V}_1\mathcal{V}_2\boldsymbol{\zeta}_{lt}^{*}-\left[X_{lt}^{*\top}\mathcal{V}_1\mathcal{V}_2X_{lt}^{*}\right]^{-1}X_{lt}^{*\top}\mathcal{V}_1$$
$$\times\left[\mathcal{V}_1+\boldsymbol{I}-\mathcal{V}_1\mathcal{V}_2X_{lt}^{*}\left(X_{lt}^{*\top}\mathcal{V}_1\mathcal{V}_2X_{lt}^{*}\right)^{-1}X_{lt}^{*\top}\mathcal{V}_1\right]^{-1}$$
$$\times\left[\mathcal{V}_1\mathcal{V}_2\boldsymbol{\zeta}_{lt}^{*}+\mathcal{V}_2\boldsymbol{\zeta}_{ll\text{-}lt}\right].$$

The $i$th element of $\boldsymbol{\zeta}_{lt}^{*}$, $i=1,\ldots,n_{ll}/2$, is,

$$\zeta_{lt(i)}=\frac{n_i}{\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)}-1.$$

The $i$th element of $\boldsymbol{\zeta}_{ll\text{-}lt}$, $i=1,\ldots,n_{ll}/2$, is,

$$\zeta_{ll\text{-}lt(i)}=\frac{t_i-n_i}{\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)}-1.$$

It is straightforward to show that $[\mathcal{V}_1\mathcal{V}_2\boldsymbol{\zeta}_{lt}^{*}+\mathcal{V}_2\boldsymbol{\zeta}_{ll\text{-}lt}]$ is, approximately, a vector of zeros. To show this, consider, without loss of generality, the $i$th element of this vector,

$$\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)\times\left[\frac{n_i}{\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)}-1\right]$$
$$+\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)\times\left[\frac{t_i-n_i}{\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)}-1\right]$$
$$=t_i-\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)\times\left[1+\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)\right].$$

Due to the Poisson approximation to the Binomial distribution,

$$\exp\left(X_{ll\text{-}lt(i)}\boldsymbol{\lambda}_{ll\text{-}lt}^{it}\right)\simeq t_i\frac{1}{1+\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)}.$$

Thus, the elements of vector $[\mathcal{V}_1\mathcal{V}_2\boldsymbol{\zeta}_{lt}+\mathcal{V}_2\boldsymbol{\zeta}_{ll\text{-}lt}]$ are all approximately zero, and the first $n_\beta$ elements of $(X_{rll}^{\top}\mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}^{it})X_{rll})^{-1}X_{rll}^{\top}\mathcal{V}_{\text{log-linear}}(\boldsymbol{\lambda}^{it})\boldsymbol{\zeta}_{\text{log-linear}}$ are approximately equal to,

$$\left(X_{lt}^{\top}\mathcal{V}_{\text{logistic}}X_{lt}\right)^{-1}X_{lt}^{*\top}\mathcal{V}_1\mathcal{V}_2\boldsymbol{\zeta}_{lt}^{*}$$
$$=\left(X_{lt}^{\top}\mathcal{V}_{\text{logistic}}X_{lt}\right)^{-1}X_{lt}^{\top}\mathcal{V}_{1,\text{reduced}}\left(\mathcal{V}_{2,1}\ldots\mathcal{V}_{2,(j_1-1)\times j_2\times\cdots\times j_q}\right)\boldsymbol{\zeta}_{lt}^{*}.$$

Using the Poisson approximation to the Binomial distribution, for the $i$th element of $\boldsymbol{\zeta}_{lt}^{*}$, and assuming without any loss of generality that $i < n_{lt}$,

$$\begin{aligned}\zeta_{lt(i)}^{*} &\simeq \frac{n_i}{\exp\left(X_{rll(i)}\boldsymbol{\lambda}_r^{it}\right)} - 1 = \frac{n_i}{\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right) t_i \frac{1}{1+\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)}} - 1 \\ &= \frac{n_i\left(1+\exp(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it})\right) - t_i\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)}{t_i\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)}.\end{aligned}$$

Thus,

$$\zeta_{lt(i)}^{*} \simeq \left(1+\exp\left(X_{lt(i)}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it}\right)\right)^{-1}\zeta_{\text{logistic}(i)}.$$

Therefore, the updating step for $\boldsymbol{T}_r\boldsymbol{\lambda}_r$ is,

$$\begin{aligned}\boldsymbol{T}_r\boldsymbol{\lambda}_r^{it+1} &= \boldsymbol{T}_r\boldsymbol{\lambda}_r^{it} + \left(X_{lt}^{\top}\mathcal{V}_{\text{logistic}}X_{lt}\right)^{-1}X_{lt}^{\top} \\ &\quad\times \mathcal{V}_{1,\text{reduced}}\left(\boldsymbol{I}+\mathcal{V}_{1,\text{reduced}}\right)^{-1}\left(\mathcal{V}_{2,1}\ldots\mathcal{V}_{2,(j_1-1)\times j_2\times\cdots\times j_q}\right)\left(\boldsymbol{\zeta}_{\text{logistic}}^{it\top}\cdots\boldsymbol{\zeta}_{\text{logistic}}^{it\top}\right)^{\top}. \\ &= \boldsymbol{T}_r\boldsymbol{\lambda}_r^{it} + \left(X_{lt}^{\top}\mathcal{V}_{\text{logistic}}X_{lt}\right)^{-1}X_{lt}^{\top} \\ &\quad\times \mathcal{V}_{1,\text{reduced}}\left(\boldsymbol{I}+\mathcal{V}_{1,\text{reduced}}\right)^{-1}\left(\mathcal{V}_{2,1}+\cdots+\mathcal{V}_{2,(j_1-1)\times j_2\times\cdots\times j_q}\right)\boldsymbol{\zeta}_{\text{logistic}}^{it}.\end{aligned}$$

If the logistic regression was to be fitted directly, obtaining the MLE would be based on the IRLS algorithm,

$$\boldsymbol{\beta}^{it+1} = \boldsymbol{\beta}^{it} + \left(X_{lt}^{\top}\mathcal{V}_{\text{logistic}}(\boldsymbol{\beta}^{it})X_{lt}\right)^{-1}X_{lt}^{\top}\times\mathcal{V}_{\text{logistic}}(\boldsymbol{\beta}^{it})\boldsymbol{\zeta}_{\text{logistic}}^{it}.$$

By replacing the sum of the elements of the $\mathcal{V}_{2,k}$ matrices with the approximate values given in (3), we observe that, asymptotically, the updating step is the same for both $\boldsymbol{T}_r\boldsymbol{\lambda}_r$ and $\boldsymbol{\beta}$. Thus, if the starting point for $\boldsymbol{T}_r\boldsymbol{\lambda}_r$ is the same as the starting point for $\boldsymbol{\beta}$, the iterative algorithm would give the same MLE for the logistic regression parameters and the corresponding log-linear model parameters. The IRLS algorithm is robust to different starting values when the likelihood is not flat. Therefore, asymptotically, $\hat{\boldsymbol{\beta}} \simeq \boldsymbol{T}_r\hat{\boldsymbol{\lambda}}_r$ and the proof is complete. □

## References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken

Bapat RB (2011) Graphs and matrices. Springer, Hindustan Book Agency, New Delhi

Consonni G, Veronese P (2008) Compatibility of prior specifications across linear models. Stat Sci 23:232–353

Dellaportas P, Forster JJ (1999) Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 86:615–633

Dellaportas P, Forster JJ, Ntzoufras I (2012) Joint specification of model space and parameter space prior distributions. Stat Sci 27:232–246

Edwards D, Havránek T (1985) A fast procedure for model search in multi-dimensional contingency tables. Biometrika 72:339–351

Fouskakis D, Ntzoufras I, Draper D (2015) Power-expected-posterior priors for variable selection in Gaussian linear models. Bayesian Anal 10:75–107

Held L, Sabanès Bovè D, Gravestock I (2015) Approximate Bayesian model selection with the deviance statistic. Stat Sci. http://www.imstat.org/sts/future_papers.html. Accessed 17 Mar 2016

Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934

Liang F, Paulo R, Molina G, Clyde MA, Berger JO (2008) Mixtures of $g$-priors for Bayesian variable selection. J Am Stat Assoc 103:410–423

Lutkepohl H (1996) Handbook of matrices. Wiley, Chichester

Mukhopadhyay M, Samantha T (2016) A mixture of $g$-priors for variable selection when the number of regressors grows with the sample size. Test. doi:10.1007/s11749-016-0516-0

Ntzoufras I, Dellaportas P, Forster JJ (2003) Bayesian variable and link determination for generalized linear models. J Stat Plan Inference 111:165–180

Ntzoufras I (2009) Bayesian modelling using WinBugs. Wiley, Hoboken

O'Hagan A (1995) Fractional Bayes factors for model comparison. J R Stat Soc Ser B 57:99–138

O'Hagan A, Forster JJ (2004) Bayesian inference, 2nd edn. vol 2B of 'Kendall's Advanced Theory of Statistics'. Arnold, London

Overstall A, King R (2014a) A default prior distribution for contingency tables with dependent factor levels. Stat Methodol 16:90–99

Overstall A, King R (2014b) Conting: an R package for Bayesian analysis of complete and incomplete contingency tables. J Stat Softw 58:1–27

Papathomas M, Richardson S (2016) Exploring dependence between categorical variables: benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms. J Stat Plan Inference 173:47–63

Papathomas M, Dellaportas P, Vasdekis VGS (2011) A novel reversible jump algorithm for generalized linear models. Biometrika 98:231–236

Rohatgi VK (1976) An introduction to probability theory and mathematical statistics. Wiley, New York

Sabanès Bovè D, Held L (2011) Hyper-g priors for generalized linear models. Bayesian Anal 6:387–410

Wang X, George GI (2007) Adaptive Bayesian criteria in variable selection for generalized linear models. Stat Sinica 17:667–690

Wood SN (2006) Generalized additive models. An introduction with R, Chapman and Hall/CRC, New York

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with $g$-prior distributions. In: Goel PK, Zellner A (eds) Bayesian inference and decision techniques: essays in honor of Bruno de Finetti. North-Holland/Elsevier, Amsterdam, pp 233–243