Bayesian (conditionally) conjugate inference for discrete data models

Jon Forster
(University of Southampton)

with Mark Grigsby (Procter and Gamble?)
Emily Webb (Institute of Cancer Research)

Table 1: Alcohol intake, hypertension and obesity (Knuiman and Speed, 1988)

Obesity	Hypertension	Alcohol intake (drinks/day) 0	1-2	3-5	> 5
Low	Yes	5	9	8	10
	No	40	36	33	24
Average	Yes	6	9	11	14
	No	33	23	35	30
High	Yes	9	12	19	19
	No	24	25	28	29

The data

Sample data consists of counts of (multivariate) categorical variables, recorded for each unit in the sample.

Units $i = 1, \ldots, n$ are classified by variables $j = 1, \ldots, p$, where variable j has m_j categories.

Categorical response vectors $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{ip})$ are observed.

The contingency table $\boldsymbol{y}$, derived from $\{\boldsymbol{y}_i, i = 1, \ldots, n\}$ has $m = \prod_1^p m_j$ cells, with cell counts

$$\boldsymbol{y} = \sum_{i=1}^{n} \bigotimes_{j=1}^{p} \boldsymbol{y}'_{ij}$$

where $[\boldsymbol{y}'_{ij}]_k = I[y_{ij} = k], k = 1, \ldots, m_j$.

The unrestricted model

For any unit i, the marginal distribution of $\boldsymbol{y}$ ($n = 1$) can be expressed in unconstrained form as

$$p(\boldsymbol{y}) = \prod_{k=1}^{m} \pi_k^{y_k} = \exp\left\{ \sum_{k=1}^{m} \theta_k y_k - n \log\left(\sum_{k=1}^{m} \exp \theta_k \right) \right\} \tag{1}$$

where $\pi = (\pi_1, \ldots, \pi_m)$ is a vector of cell probabilities and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is a corresponding multivariate logit

$$\theta_k = \log \pi_k - \sum_{\ell=1}^{m} a_\ell \log \pi_\ell$$

where $\sum a_\ell = 1$, $a_\ell \geq 0$. Typically $a_\ell = I[\ell = 1]$ or $a_\ell = I[\ell = m]$ (reference cell logit) though $a_\ell = 1/m$ (centred logit) can sometimes be more useful.

Typically, we assume that individual classifications $\boldsymbol{y}_i$ are conditionally independent given a common $\boldsymbol{\theta}$, in which case (1) holds for $n > 1$. Distribution of $\boldsymbol{y}$ is a multivariate natural exponential family.

Conjugate inference for the unrestricted model

A Bayesian conjugate prior for cell probabilities π of the unrestricted model is Dirichlet($\boldsymbol{\alpha}$), that is

$$p(\pi) \propto \prod_{k=1}^{m} \pi_k^{\alpha_k - 1}$$

or equivalently, for $\boldsymbol{\theta}$,

$$p(\boldsymbol{\theta}) \propto \exp\left\{ \sum_{k=1}^{m} \theta_k \alpha_k - \alpha_+ \log\left(\sum_{k=1}^{m} \exp \theta_k \right) \right\} \tag{2}$$

Correspondence of (2) and (1) makes this the Diaconis-Ylvisaker conjugate prior.

[Alternatives include priors based on multivariate normal distribution for θ; some advantages, some drawbacks]

Incorporating structure with parsimonious models

General log-linear models

$$\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$$

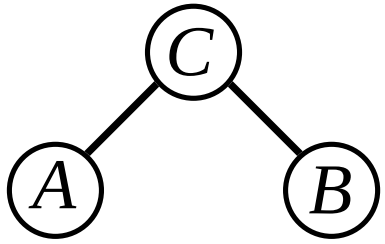
for some $n \times r$ $\boldsymbol{X}$ satisfying the constraint on $\boldsymbol{\theta}$

Log-linear interaction models for contingency tables

Imply specific forms for $\boldsymbol{X}$ (with or without hierarchy constraints)

(Undirected) graphical models

Hierarchical models specified purely by conditional independence properties (constraints)



Conjugate inference for the general log-linear model

$$p(\boldsymbol{y}) \propto \exp\left\{\sum_{k=1}^{r} \beta_k t_k - n \log\left(\sum_{k=1}^{m} \exp[\boldsymbol{X}\boldsymbol{\beta}]_k\right)\right\} \tag{3}$$

where $\boldsymbol{t} = \boldsymbol{X}^T\boldsymbol{y}$.

The Diaconis-Ylvisaker conjugate prior for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}) \propto \exp\left\{\sum_{k=1}^{r} \beta_k s_k - \alpha \log\left(\sum_{k=1}^{m} \exp[\boldsymbol{X}\boldsymbol{\beta}]_k\right)\right\} \tag{4}$$

where $\boldsymbol{s}$ and α are hyperparameters.

Massam, Liu and Dobra (2008) investigate this structure in detail for hierarchical log-linear models, and provide lots of interesting results.

Forthcoming attraction

Bayesian structural learning and estimation in Gaussian graphical models and hierarchical log-linear models

starring Adrian Dobra

In theatre(s) from July 8

Conditional Dirichlet distributions and compatibility (Grigsby, 2001)

Suppose we derive a prior for a general log-linear model by taking a Dirichlet for the unconstrained model and conditioning (*on the log-linear scale*).

Then, essentially trivially, we arrive at

$$p(\boldsymbol{\beta}) \propto \exp\left\{\sum_{k=1}^{r} \beta_k s_k - \alpha \log\left(\sum_{k=1}^{m} \exp[\boldsymbol{X}\boldsymbol{\beta}]_k\right)\right\} \tag{4}$$

as before.

Hence, these conjugate priors are compatible (in a conditional sense; see Dawid and Lauritzen, 2001).

Conditioning on the linear ($\boldsymbol{\theta}$) scale ensures here that a proper initial prior leads to a proper prior on the submodel. See also Gûnel and Dickey (1974).

Hyper-Dirichlet equivalence

MLD, Leucari, Grigsby show that, for a decomposable undirected graphical model, the conjugate prior (4) is equivalent to a hyper-Dirichlet (Dawid and Lauritzen, 1993).

Grigsby proof

Based on a perfect ordering, and standard decomposition

$$p(\boldsymbol{y}_i) = \prod_{j=1}^{p} p_j(y_{ij}|y_{i,\mathrm{pa}(j)})$$

where $\pi_j(\ell_{\mathrm{pa}(j)}) = \{p_j(k|\ell_{\mathrm{pa}(j)}), k = 1, \ldots m_j\}$ is a m_j-vector of probabilities for each distinct combination $\ell_{\mathrm{pa}(j)}$ of levels of $\mathrm{pa}(j)$.

With this parameterisation, natural (closed under sampling) prior family is

$$\pi_j(\ell_{pa(j)}) \sim \text{Dirichlet}(\boldsymbol{\alpha}_{j,\ell_{pa(j)}}) \tag{5}$$

independently, for each variable j and each parent combination $\ell_{pa(j)}$.

This is the product Dirichlet described by Cowell et al (1999), which here is equivalent to the hyper-Dirichlet for particular choices of $\{\boldsymbol{\alpha}_{j,\ell_{pa(j)}}\}$ (directed hyper-Dirichlet).

Use this distribution within the (conditional) logit parameterisation

$$\theta_j(k, \ell_{pa(j)}) = \log p_j(k, |\ell_{pa(j)}) - \log p_j(1|\ell_{\text{pa}(j)})$$

Then the (directed) hyper-Dirichlet and conditional Dirichlet densities are functionally equivalent ('likelihood-like').

Remains to prove that Jacobian = 1; achieved via identifying a correspondence between $\{\theta_j(k, \ell_{pa(j)})\}$ and log-linear parameters $\boldsymbol{\beta}$ for which derivative matrix is upper triangular with unit diagonal.

'Non-hyper' product Dirichlet priors for perfect DAG models

Is there any role for using independent

$$\pi_j(\ell_{pa(j)}) \sim \text{Dirichlet}(\boldsymbol{\alpha}_{j,\ell_{pa(j)}}) \tag{5}$$

with non-consistent $\{\boldsymbol{\alpha}_{j,\ell_{pa(j)}}\}$?

Potentially, unattractive as typically such a prior family will not be closed under changes of (non-trivial) decomposition. For example for a pair of binary variables

$$\perp\!\!\!\perp\{\pi_1 \sim \text{Beta}(\boldsymbol{\alpha}_1), \pi_2(1) \sim \text{Beta}(\boldsymbol{\alpha}_{21}), \pi_2(2) \sim \text{Beta}(\boldsymbol{\alpha}_{22})\}$$
$$\not\Rightarrow \perp\!\!\!\perp\{\pi_2 \sim \text{Beta}(\boldsymbol{\alpha}_2), \pi_1(1) \sim \text{Beta}(\boldsymbol{\alpha}_{11}), \pi_1(2) \sim \text{Beta}(\boldsymbol{\alpha}_{12})\}$$

However, for certain decomposable models, for particular decompositions, the Jeffreys prior has this form.

Jeffreys prior

Recall Jeffreys rule for constructing a default prior distribution

$$f(\pi) \propto |\mathcal{I}(\pi)|^{1/2},$$

invariant under reparameterisation.

For a decomposable undirected graphical model, parameterised via a perfect DAG and $\{\pi_j(\ell_{pa(j)})\}$, the Jeffreys prior is

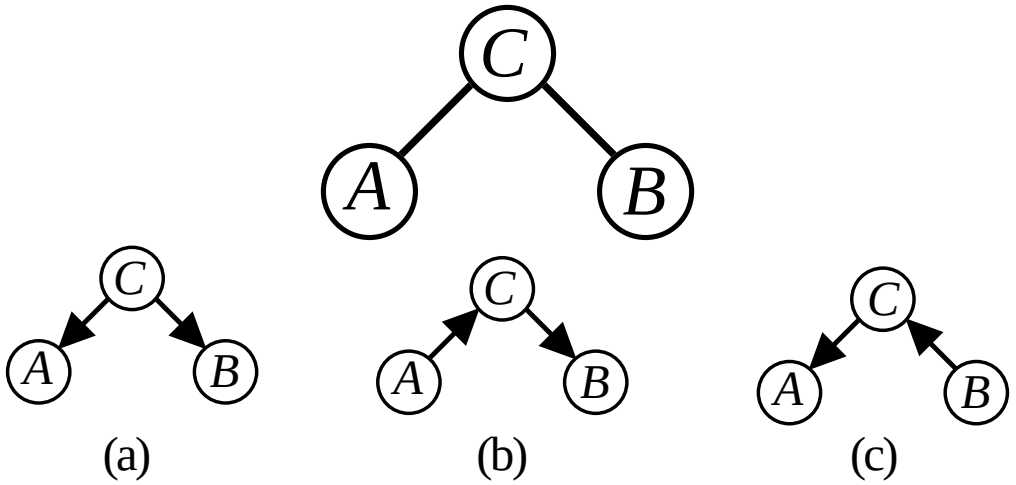
$$f(\pi) \propto \left(\prod_j \prod_{\ell_{\mathrm{pa}(j)}} \left[\sum_{\ell_{\overline{\mathrm{pa}}(j)}} \prod_{j'} p_{j'}(\ell_{j'}|\ell_{\mathrm{pa}(j')}) \right]^{\frac{m_j-1}{2}} \right) \prod_j \prod_{\ell_{\{j\}\cup\mathrm{pa}(j)}} p_j(\ell_j|\ell_{\mathrm{pa}(j)})^{-\frac{1}{2}} \tag{6}$$

Jeffreys prior simplifies to product Dirichlet, when the summation in (6) can be simplified to a single product term for every $j, \ell_{\mathrm{pa}(j)}$.

Jeffreys prior simplification

The Jeffreys prior simplifies to product Dirichlet for any 'grandparent-free' perfect ordering.

Such an ordering is only possible for decomposable graphs whose junction tree has a common separator between nodes.



For (a) the Jeffreys prior simplifies to independent symmetric Dirichlet($\alpha\mathbf{1}$) distributions with

$$\alpha_C = (m_A + m_B - 3)/2, \quad \alpha_{A,\ell_C} = \alpha_{B,\ell_C} = 1/2 \text{ for all } \ell_C$$

Inference and computation

For Dirichlet-equivalent priors, parameteric inference and marginal likelihood computation (for model determination) are easy.

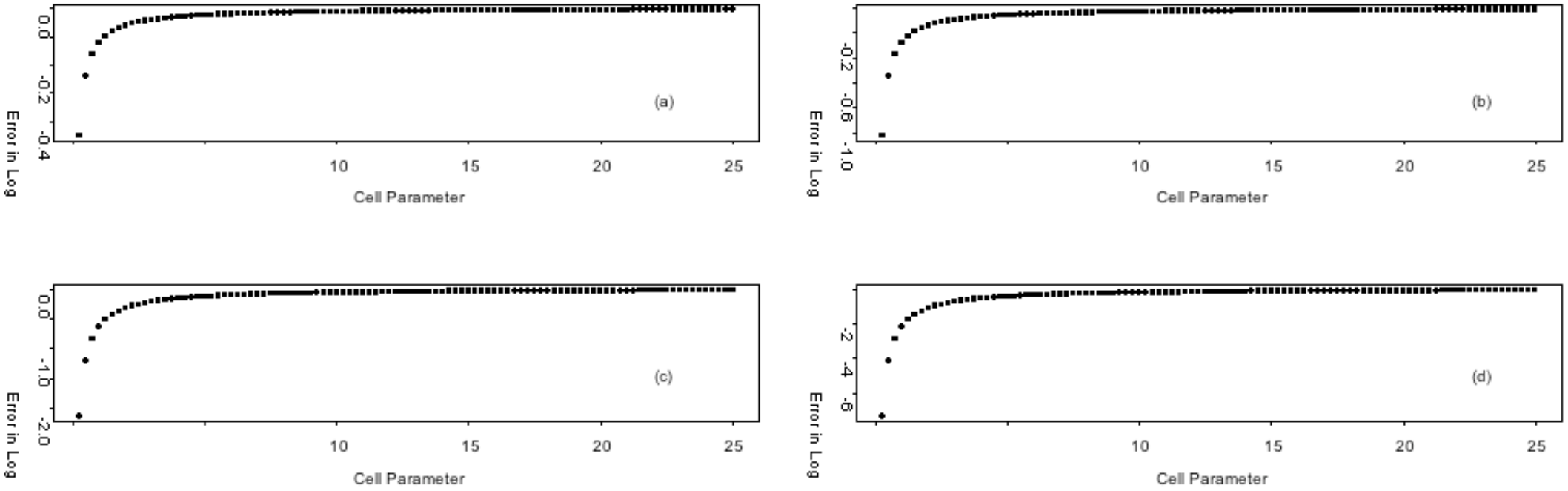
For other priors we have discussed, numerical methods are required, and for non-decomposable conjugates and non-Dirichlet Jeffreys, prior normalising constants must be computed for marginal likelihoods.

(Markov chain) Monte Carlo methods or numerical integration based on Laplace's method work well.

For conditional (generalized hyper) Dirichlet for nondecomposable models, Laplace can be computed simply in standard software with log-linear model fitting, and basic matrix functionality.

However, beware of Laplace's method for prior normalising constant.

Convergence of Laplace's method for prior normalising constant



(a) A + B [2]
(b) AB+BC [5]
(c) ABC [8]
(d) ABC [27]

Can be improved by bridge sampling from prior and a normal 'approximation' (time-consuming, but can build a catalogue)

Alternative computation?

For model determination and conditional Dirichlet priors, it may be possible to exploit the Savage-Dickey density ratio.

Recall that, for comparing models $M_1 : y \sim p_1(y|\theta, \phi)$ and $M_0 : y \sim p_0(y|\phi)$ where $p_0(y|\phi) = p_1(y|\theta_0, \phi)$, we have the Bayes factor

$$\frac{p_0(y)}{p_1(y)} = \frac{p_1(\theta_0|y)}{p_1(\theta_0)}$$

provided that

$$p_0(\phi) = p_1(\phi|\theta_0).$$

This constraint is exactly the same as used in deriving the conditional Dirichlet, so for example models ABC and AB+AC+BC can be compared by generating the missing interaction parameter (log-contrast) under the Dirichlet prior and posterior for ABC, and calculating the ratio of two density estimates at zero.

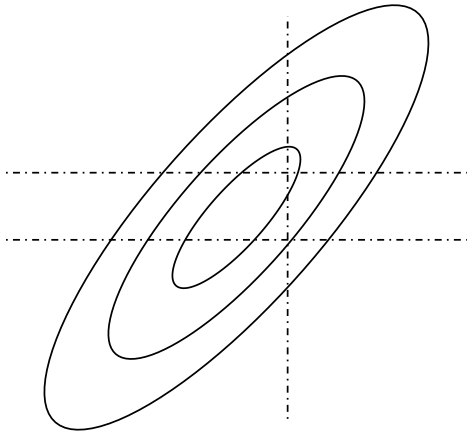
How feasible is this generally?

Posterior model probabilities for Table 1

Model	Posterior probability	Posterior probability (ordinal)
$OH + AH$	0.036	0.725
$A + OH$	0.643	0.091
AOH	0.000	0.084
$OH + OA$	0.000	0.053
$O + AH$	0.017	0.027
$OA + AH$	0.000	0.013
$O + A + H$	0.304	0.005
$H + OA$	0.000	0.002

Ordinal probit models – multivariate (Chib and Greenberg, 1998)

$\boldsymbol{z}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_m)$ is a latent continuous variable
$y_{ij} = c$ if $\alpha_{j,c-1} < z_i \leq \alpha_{j,c}$, $\qquad (\alpha_{j,0} = -\infty,\ \alpha_{j,m_j} = \infty)$



Prior is
$\boldsymbol{\beta} \sim$ MV Normal
$\alpha_{j,c} \sim$ Indept Uniform subject to ordering constraint
$\boldsymbol{\Sigma}_m \sim$ distribution consistent with any constraints

Identifiability constraints – $\sigma_{ii} = 1,\ \alpha_{j,1} = 0,\ j = 1, \ldots, p.$

An alternative parameterisation

Constrain $\alpha_{j,1} = -\alpha_{j,m_j-1} = \Phi^{-1}\left(\frac{1}{m_j}\right),\ j = 1, \ldots, p.$

Then $\boldsymbol{\Sigma}$ is unconstrained and can be given a (hyper) Inverse Wishart prior. Conditionals are then straightforward to sample.

Not possible if any $k_i = 2$ (binary variable).

Instead, consider the Cholesky decomposition

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Phi}^T \boldsymbol{\Phi}$$

where $\boldsymbol{\Phi}$ is upper triangular.

The elements of $\boldsymbol{\Phi}$ appear in the decomposition

$$z_{ip} \sim \beta_p + N\left(0, \frac{1}{\phi_{pp}^2}\right)$$

$$z_{i,p-1}|z_{ip} \sim \beta_{p-1} - \frac{\phi_{p-1,p}}{\phi_{p-1,p-1}}(z_{ip} - \beta_p) + N\left(0, \frac{1}{\phi_{p-1,p-1}^2}\right)$$

$$\vdots \qquad \vdots$$

$$z_{i1}|z_{i2}, \ldots, z_{ip} \sim \beta_1 - \frac{\phi_{1p}}{\phi_{11}}(z_{ip} - \beta_p) - \ldots - \frac{\phi_{12}}{\phi_{11}}(z_{i2} - \beta_2) + N\left(0, \frac{1}{\phi_{11}^2}\right)$$

For binary (and other) variables we can constrain $\lambda_j \equiv \phi_{jj}^{-1} = 1$.

remaining $\lambda_j \equiv \phi_{jj}^{-1} \sim$ Gamma
$\boldsymbol{\psi}_j \equiv (\phi_{j,j+1}, \ldots, \phi_{jp})\phi_{jj}^{-1} \,|\, \phi_{jj} \sim$ MV Normal

[Equivalence with (hyper) inverse Wishart; Roverato (2002)]

Graphical models

Gaussian DAG models for $\boldsymbol{z}$ ('graphical' ordinal probit models for $\boldsymbol{y}$) can be specified by setting certain $\psi_{jk} = 0$, for an appropriate ordering.

Undirected graphical models can be specified using an equivalent DAG

Conditional conjugacy allows straightforward MCMC computation

Model determination for DAG models given an ordering uses Reversible Jump MCMC with transitions between models which differ by a single edge (see also Fronk, 2002)

Model determination for undirected graphical models requires order switching.

Propose to transpose two neighbouring variables in the current ordering, with associated deterministic parameter transformation (RJMCMC allows this)

Prior must compensate for the fact that not all models are available under the same number of orderings (Order counting in the 'model jump' step; Chandran et al, 2003).

Table 2: Colouring of blackbirds (Anderson and Pemberton, 1985)

Lower Mandible	Upper Mandible	Orbital Ring 1	2	3
	1	40	19	0
1	2	0	0	0
	3	0	1	0
	1	1	6	0
2	2	1	2	1
	3	0	1	0
	1	1	2	0
3	2	0	1	1
	3	0	6	7

1 = mostly black, 2 = intermediate, 3 = mostly yellow

Predictive logarithmic scores for Table 2

Conditional independence structure	Ordinal models	Non-ordinal models
None	−178.4	−197.7
$L \perp\!\!\!\perp O \mid U$	−177.8	−186.3
$U \perp\!\!\!\perp O \mid L$	−178.3	−188.2
model-averaged	−178.4	−190.7

$$S = \sum_{i=1}^{90} \log p(\boldsymbol{y}_i | \boldsymbol{y}_{\setminus i})$$

where $\boldsymbol{y}_{\setminus i}$ represents the data $\boldsymbol{y}$ with $\boldsymbol{y}_i$ removed

Posterior model probabilities are 0.279 (unstructured), 0.427 ($L \perp\!\!\!\perp O \mid U$) and 0.293 ($U \perp\!\!\!\perp O \mid L$).