

# Bayesian model averaging for categorical data

Jon Forster

University of Southampton

(includes joint work with Emily Webb)

1. Motivating examples
2. Bayesian inference for contingency tables
3. Example: Disclosure risk estimation
4. Extra structure: Ordinal data
5. Extra structure: Square tables

Table 1: 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

Table 2: Anderson and Pemberton (1985)

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

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

Table 3: Inter-Ethnic Unions in Great Britain, 1991 (1% SAR)

Ethnic group of male partner		Ethnic group of female partner									
		1	2	3	4	5	6	7	8	9	10
1	White	126150	102	41	63	71	10	0	79	148	139
2	Black-Caribbean	225	599	8	10	4	2	0	2	3	12
3	Black-African	48	16	208	4	2	1	0	0	0	2
4	Black Other	76	3	2	62	1	0	0	0	2	1
5	Indian	134	2	4	1	1762	18	0	5	4	5
6	Pakistani	42	0	0	1	6	775	0	0	4	3
7	Bangladeshi	7	0	2	0	4	1	217	0	0	2
8	Chinese	34	0	0	0	2	0	0	234	0	0
9	Other Asian	55	4	1	1	4	4	1	2	296	6
10	Other	218	2	1	2	7	4	0	2	5	191

Table 4: Distribution of marriages in the village of Nemgéné by lineage of each spouse (Cazes, 1990)

Husband's lineage	Wife's lineage								
	Iariwa Ger.	Segiwa	Suraba	Pussuwoï	Iariwa	Tengo <sub>2</sub>	Tengo	Other	
Iariwa Ger.	<b>0</b>	2	3	3	3	4	1	3	
Segiwa	2	<b>0</b>	2	5	1	1	6	3	
Suraba	0	2	<b>0</b>	2	3	6	7	6	
Pussuwoï	1	4	4	<b>0</b>	3	5	9	7	
Iariwa	2	2	3	4	<b>1</b>	1	13	16	
Tengo <sub>2</sub>	4	0	4	8	4	<b>1</b>	16	14	
Tengo	4	6	8	10	8	14	<b>12</b>	16	
Other	2	6	7	4	17	6	21	<b>17</b>	

Table 5: Six potential key variables from the 3% 2001 Individual SAR

Restricted to 154295 individuals living in South West England

Sex (2 categories)

Age (coded into 11 categories)

Accommodation type (8 categories)

Number of cars owned or available for use (5 categories)

Occupation type (11 categories)

Family type (10 categories)

The full table has 96800 cells of which 3796 are uniques.

This is our 'population', from which we took a 3% subsample.

## Example: Disclosure risk assessment (2)

**Sample data** contains 4761 individuals in 2330 cells.

1543 (32%) are uniques, of which 114 (7%) are population uniques. Average population total in a sample unique cell is 17.

		Population								
		0	1	2	3	4	5-9	10-19	20+	Total
Sample	0	84867	3682	1694	967	631	1482	757	390	94470
	1	—	<b>114</b>	<b>110</b>	<b>118</b>	<b>104</b>	<b>313</b>	<b>322</b>	<b>462</b>	<b>1543</b>
	2	—	—	0	2	5	28	67	266	368
	3	—	—	—	0	0	1	15	140	156
	4	—	—	—	—	0	0	0	76	76
	5-9	—	—	—	—	—	0	0	125	125
	10-19	—	—	—	—	—	—	0	48	48
	20+	—	—	—	—	—	—	—	14	14
	Total	84867	3796	1804	1087	740	1824	1161	1521	96800

Sample data consists of values of categorical variables, recorded for each individual in the sample, expressed as a multiway contingency table.

The observed data are  $\mathbf{y}$  which can be thought of as

$\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ , the collection of unit level data for the sample, or as

$\mathbf{y} = (y_1, \dots, y_K)$ , the cell counts in the corresponding contingency table.

$\mathbf{Y} = (Y_1, \dots, Y_K)$  are the corresponding population cell counts.

$n$  and  $N$  are the sample and population totals respectively.

For a single model, data  $\mathbf{y}$ , model parameters  $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})}$$

Posterior  $\propto$  likelihood  $\times$  prior

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}) \, d\boldsymbol{\beta}$$

Prediction of a future observation  $y^*$  proceeds naturally through

$$p(y^*|\mathbf{y}) = \int p(y^*|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{y}) \, d\boldsymbol{\beta}$$

assuming  $y^*$  is independent of  $\mathbf{y}$  given  $\boldsymbol{\beta}$ .

Population cell frequencies  $\mathbf{Y} = (Y_1, \dots, Y_K)$  are the main parameters (unknowns).

A two stage prior distribution,  $\mathbf{Y} \sim p(\mathbf{Y}|\boldsymbol{\beta})$ ,  $\boldsymbol{\beta} \sim p(\boldsymbol{\beta})$

This implies a model for  $p(\mathbf{y}|\boldsymbol{\beta})$ , from which we get

$$p(\boldsymbol{\beta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y})} \quad \text{as before.}$$

Then, for the unknown part of the population, (assuming  $\mathbf{Y} - \mathbf{y}$  is independent of  $\mathbf{y}$ , given  $\boldsymbol{\beta}$ ) the inference is provided by the predictive

$$p(\mathbf{Y} - \mathbf{y}|\mathbf{y}) = \int p(\mathbf{Y} - \mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{y}) \, d\boldsymbol{\beta}.$$

(Ericson, 1969)

$Y$  has a multinomial( $N, \boldsymbol{\pi}$ ) distribution.

$\mathbf{y}$  has a multinomial( $n, \boldsymbol{\pi}$ ) distribution

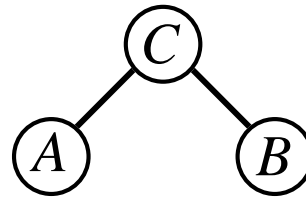
$$p(\mathbf{y}|\boldsymbol{\pi}) \propto \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k} = \prod_{i=1}^k \pi_i^{y_i}$$

$\boldsymbol{\pi}$  might have a Dirichlet prior distribution – multcategory generalisation of beta distribution.

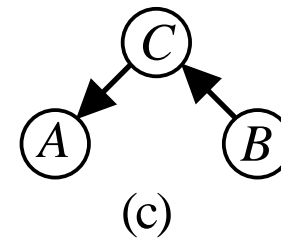
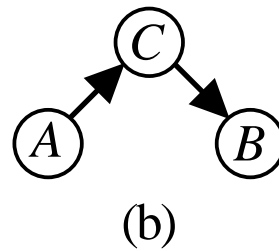
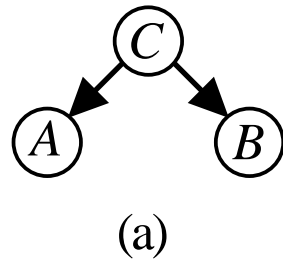
$$p(\boldsymbol{\pi}) = \frac{\Gamma(\lambda_1 + \lambda_2 + \cdots + \lambda_k)}{\Gamma(\lambda_1)\Gamma(\lambda_2)\cdots\Gamma(\lambda_k)} \pi_1^{\lambda_1-1} \pi_2^{\lambda_2-1} \cdots \pi_k^{\lambda_k-1}$$

However, we usually prefer to model  $\boldsymbol{\pi}$  as  $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\beta})$  and put a prior on  $\boldsymbol{\beta}$ .

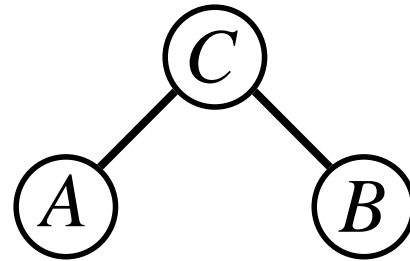
Undirected graphical models



Directed graphical models



General log-linear models



$A$  is independent of  $B$  given  $C$ , so that

$$P(A = i \text{ and } B = j | C = k) = P(A = i | C = k)P(B = j | C = k)$$

$$\Rightarrow P(A = i \text{ and } B = j \text{ and } C = k) = P(A = i | C = k)P(B = j | C = k)P(C = k)$$

or

$$\pi_{ijk} = \beta_{i|k}^A \beta_{j|k}^B \beta_k^C$$

Independent Dirichlet priors for  $P(A|C = k)$ ,  $P(B|C = k)$ , for each  $k$ , and for  $P(C)$ . Hyper-Dirichlet (Dawid and Lauritzen, 1993) is special case.

Posterior computation is generally straightforward. (Predictive?)

Allows model uncertainty to be coherently incorporated.

Full prior distribution consists of the multinomial  $p(\mathbf{Y} | \boldsymbol{\beta}_m, m)$ ,  $p(\boldsymbol{\beta}_m | m)$  the prior for  $\boldsymbol{\beta}_m$  for each  $m \in M$  and  $p(m)$ , a discrete prior distribution over the set of possible models  $M$ .

Under model uncertainty

$$p(\mathbf{Y} - \mathbf{y} | \mathbf{y}) = \sum_{m \in M} p(m | \mathbf{y}) \int p(\mathbf{Y} - \mathbf{y} | N - n, \boldsymbol{\beta}_m, m) p(\boldsymbol{\beta}_m | \mathbf{y}, m) d\boldsymbol{\beta}_m.$$

where by Bayes theorem

$$p(m | \mathbf{y}) = \frac{p(m)p(\mathbf{y} | m)}{\sum_{m \in M} p(m)p(\mathbf{y} | m)}$$

and  $p(\mathbf{y} | m) = \int p(\mathbf{y} | m, \boldsymbol{\beta}_m) p(\boldsymbol{\beta}_m | m) d\boldsymbol{\beta}_m.$

$$p(\mathbf{Y} - \mathbf{y}|\mathbf{y}) = \sum_m p(m|\mathbf{y}) \int p(\mathbf{Y} - \mathbf{y}|N - n, \boldsymbol{\beta}_m, m) p(\boldsymbol{\beta}_m|\mathbf{y}, m) d\boldsymbol{\beta}_m.$$

is a weighted average of the posterior distributions under the various models. The posterior model probabilities may not be of interest in themselves – interpret them as weights for prediction.

Posterior predictive expectations of any function of  $\mathbf{Y}$  will also be a model average

$$E[g(\mathbf{Y})|\mathbf{y}] = \sum_m p(m|\mathbf{y}) E[g(\mathbf{Y})|\mathbf{y}, m].$$

Modelling provides ‘structured smoothing’.

Table 3: Six potential key variables from the 3% 2001 Individual SAR

Restricted to 154295 individuals living in South West England

Sex (2 categories)

Age (coded into 11 categories)

Accommodation type (8 categories)

Number of cars owned or available for use (5 categories)

Occupation type (11 categories)

Family type (10 categories)

The full table has 96800 cells of which 3796 are uniques.

This is our 'population', from which we took a 3% subsample.

**Sample data** contains 4761 individuals in 2330 cells.

1543 (32%) are uniques, of which 114 (7%) are population uniques. Average population total in a sample unique cell is 17.

		Population								
		0	1	2	3	4	5-9	10-19	20+	Total
Sample	0	84867	3682	1694	967	631	1482	757	390	94470
	1	—	<b>114</b>	<b>110</b>	<b>118</b>	<b>104</b>	<b>313</b>	<b>322</b>	<b>462</b>	<b>1543</b>
	2	—	—	0	2	5	28	67	266	368
	3	—	—	—	0	0	1	15	140	156
	4	—	—	—	—	0	0	0	76	76
	5-9	—	—	—	—	—	0	0	125	125
	10-19	—	—	—	—	—	—	0	48	48
	20+	—	—	—	—	—	—	—	14	14
	Total	84867	3796	1804	1087	740	1824	1161	1521	96800

Either population uniqueness, or a sample unique

$$I[Y_i = 1, y_i = 1]$$

or Benedetti-Franconi

$$\frac{1}{Y_i}$$

As is true for most such measures, these are functions of  $\mathbf{Y}$  and can be thought of as probabilities  $P(\text{event}|\mathbf{Y})$ , given the population  $\mathbf{Y}$ .

Corresponding Bayesian predictive probabilities are posterior expectations

$$P(\text{event}|\mathbf{y}) = E[P(\text{event}|\mathbf{Y})|\mathbf{y}]$$

so the above measures are estimated by the predictive probabilities

$$p(Y_i = 1|\mathbf{y}) = p(Y_i - y_i = 0|\mathbf{y}) \quad \text{for sample uniques}$$

and

$$E(1/Y_i|\mathbf{y}) = \sum_{j=0}^{N-n} \frac{1}{j + y_i} p(Y_i - y_i = j|\mathbf{y}).$$

## Computational difficulties

1. Evaluating integrals – may be mathematically intractable
2. Number of models is large.
3. Number of possible values of (multivariate)  $\mathbf{Y}$  is large.

Monte Carlo methods of computation are possible but time-consuming.

Approximating the posterior distribution of  $\pi_i$  by a gamma( $\alpha, \beta$ ) distribution, with correct mean and variance (straightforward to evaluate for hyper-Dirichlet) gives

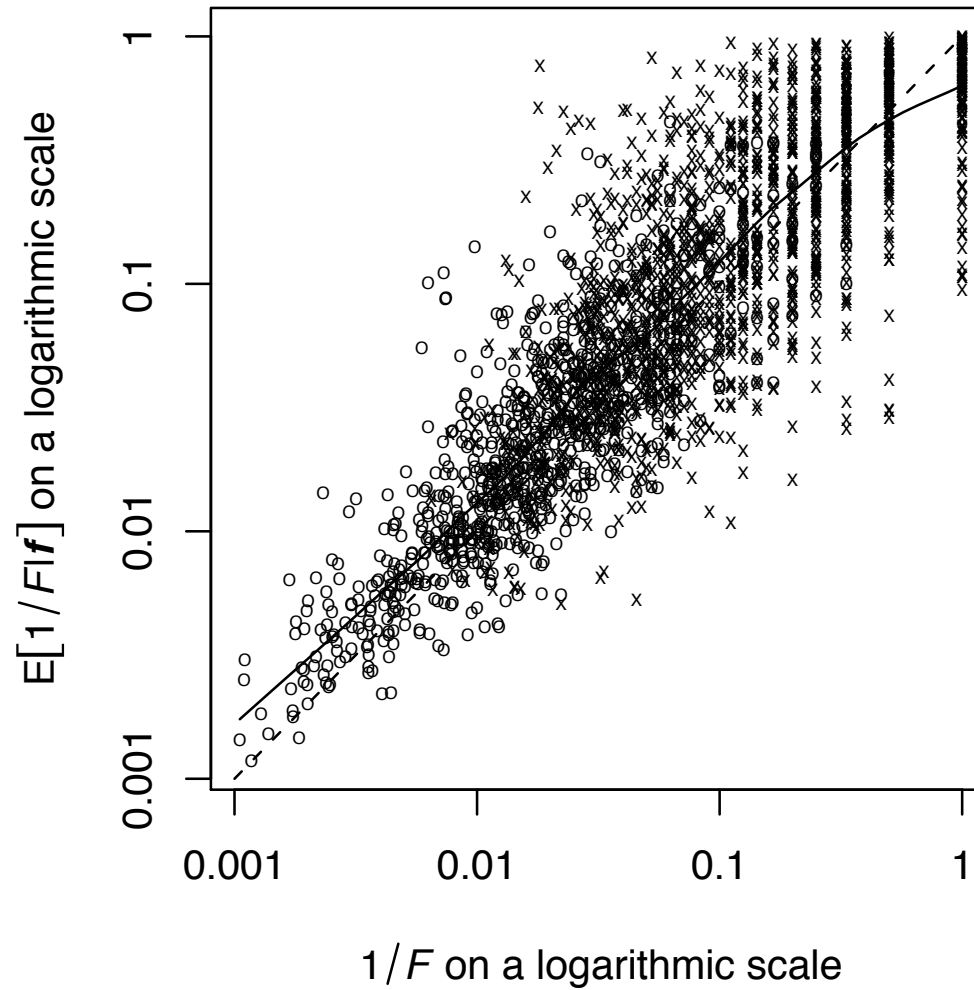
$$P(Y_i = y_i | \mathbf{y}) \approx \left( \frac{\beta}{\beta + u} \right)^\alpha$$

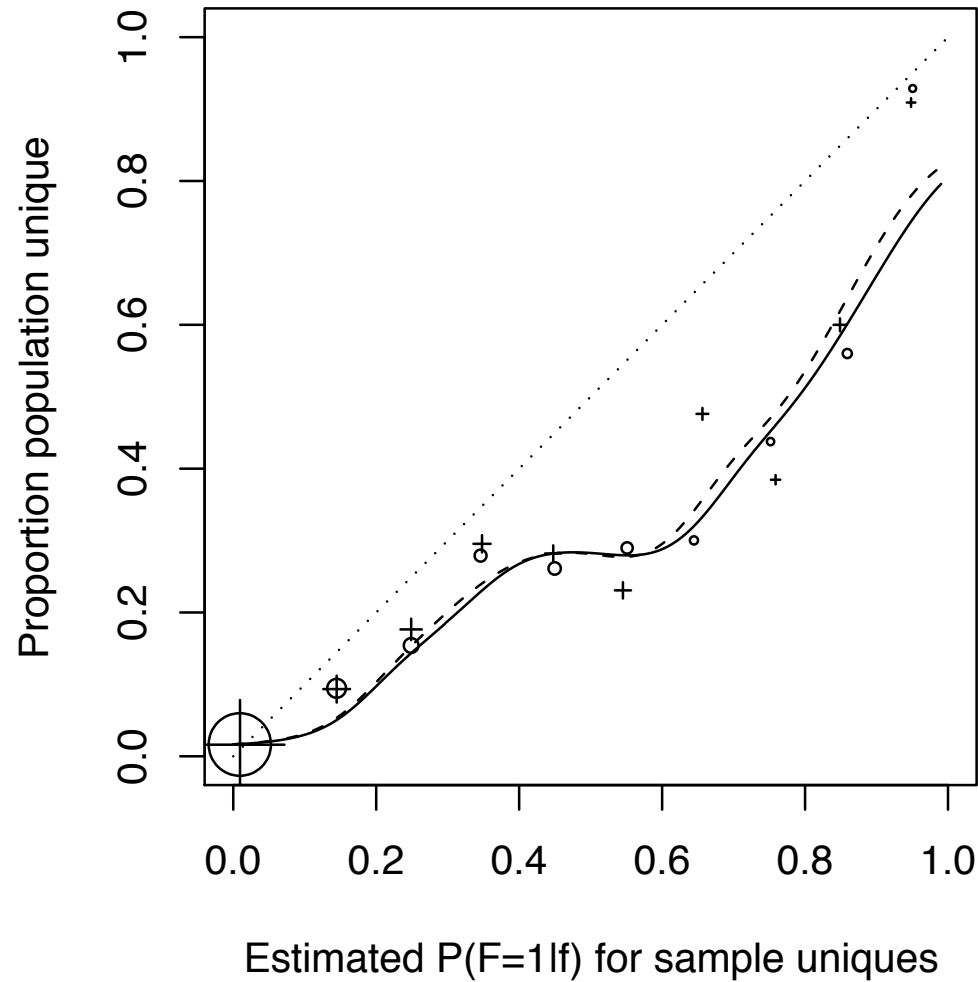
and

$$E[1/Y_i | \mathbf{y}] \approx \left( \frac{\beta}{\beta + u} \right)^\alpha \frac{1}{y_i} {}_2F_1(\alpha, y_i, y_i + 1, u/(\beta + u)).$$

which are fast to evaluate and accurate.

Also fast search strategy for identifying high probability models.





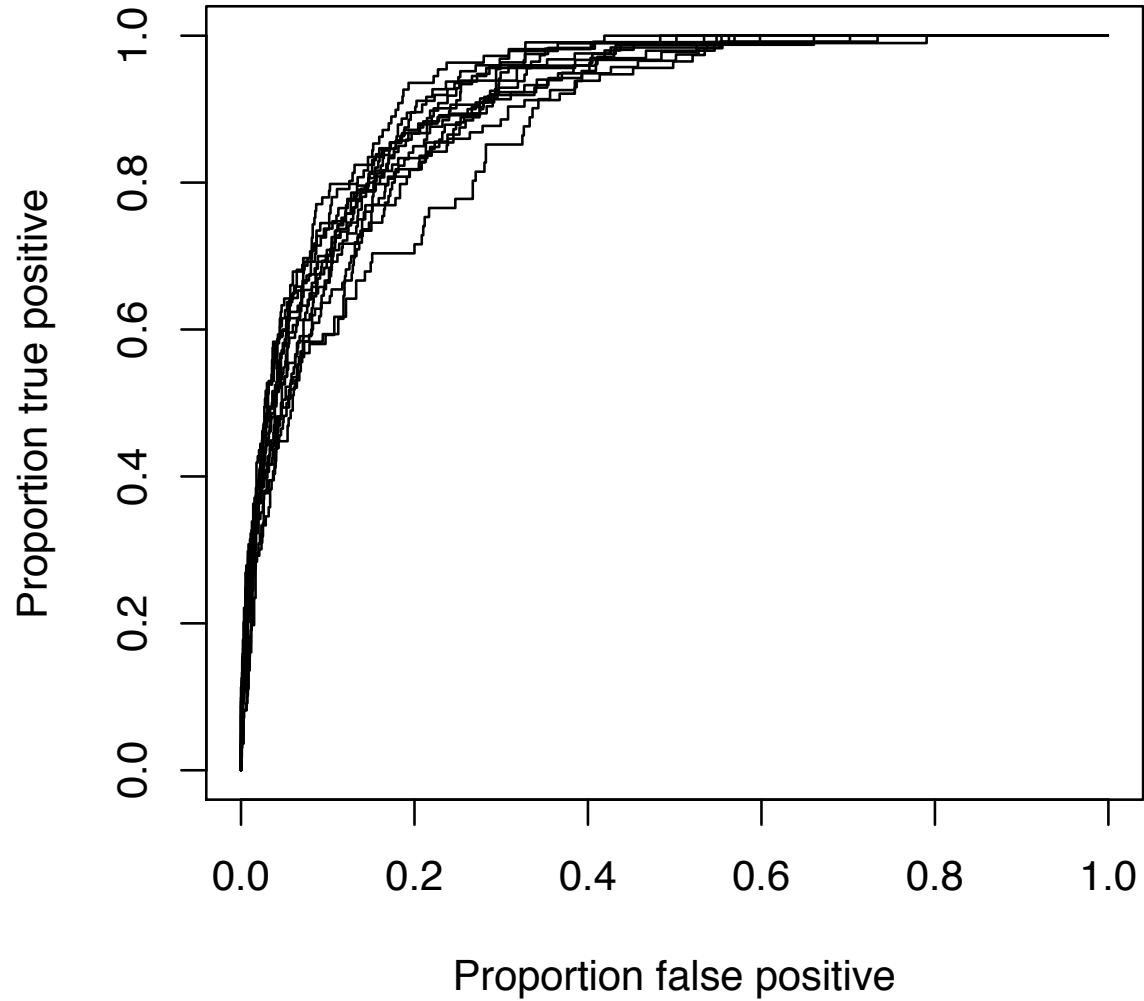


Table 1: Knuiman and Speed (1988)

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

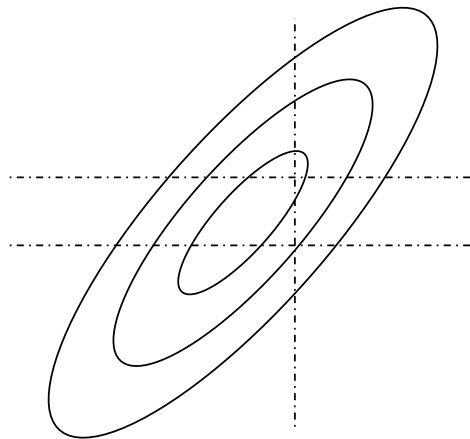
Posterior model probabilities for Table 1

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

Chib and Greenberg (1998), generalising Albert and Chib (1993)

$z_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_m)$  is a latent continuous variable

$y_{ij} = c$  if  $\alpha_{j,c-1} < z_i \leq \alpha_{j,c}$ ,  $(\alpha_{j,0} = -\infty, \alpha_{j,m_j} = \infty)$



Prior is

$\boldsymbol{\mu} \sim$  MV Normal

$\alpha_{j,c} \sim$  Inddept Uniform subject to ordering constraint

$\boldsymbol{\Sigma}_m \sim$  distribution consistent with any constraints

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

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

Then  $\Sigma$  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

$$\Sigma^{-1} = \mathbf{P}^T \mathbf{P}$$

where  $\mathbf{P}$  is upper triangular.

[Daniels and Pourahmadi, 2002, Smith and Kohn, 2002]

The elements of  $\mathbf{P}$  appear in the decomposition

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

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

$$\vdots \quad \quad \quad \vdots$$

$$z_{i1}|z_{i2}, \dots, z_{ip} \sim \mu_1 - \frac{\phi_{1p}}{\phi_{11}}(z_{ip} - \mu_p) - \dots - \frac{\phi_{12}}{\phi_{11}}(z_{i2} - \mu_2) + N\left(0, \frac{1}{\phi_{11}^2}\right)$$

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

remaining  $\lambda_j \equiv \phi_{jj}^{-1} \sim \text{Gamma}$

$\boldsymbol{\psi}_j \equiv (\phi_{j,j+1}, \dots, \phi_{jp})\phi_{jj}^{-1} | \phi_{jj} \sim \text{MV Normal}$

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

Gaussian DAG models for  $\mathbf{z}$  ('graphical' ordinal probit models for  $\mathbf{y}$ ) 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 (RJCMC 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	1	40	19	0
	2	0	0	0
	3	0	1	0
2	1	1	6	0
	2	1	2	1
	3	0	1	0
3	1	1	2	0
	2	0	1	1
	3	0	6	7

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

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

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

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

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

Table 3: Inter-Ethnic Unions in Great Britain, 1991 (1% SAR)

Ethnic group of male partner		Ethnic group of female partner									
		1	2	3	4	5	6	7	8	9	10
1	White	126150	102	41	63	71	10	0	79	148	139
2	Black-Caribbean	225	599	8	10	4	2	0	2	3	12
3	Black-African	48	16	208	4	2	1	0	0	0	2
4	Black Other	76	3	2	62	1	0	0	0	2	1
5	Indian	134	2	4	1	1762	18	0	5	4	5
6	Pakistani	42	0	0	1	6	775	0	0	4	3
7	Bangladeshi	7	0	2	0	4	1	217	0	0	2
8	Chinese	34	0	0	0	2	0	0	234	0	0
9	Other Asian	55	4	1	1	4	4	1	2	296	6
10	Other	218	2	1	2	7	4	0	2	5	191

Consider the saturated log-linear model for  $\pi$  written as

$$\log \pi_{ij} = \begin{cases} \mu + \alpha_i + \alpha_j + \beta_i - \beta_j + \gamma_{ij} + \lambda_{ij} & i < j \\ \mu + \alpha_i + \alpha_j + \beta_i - \beta_j + \gamma_{ji} & i > j \\ \mu + \nu + \delta_i & i = j \end{cases}$$

Simpler models can be obtained by setting one or more of  $\nu$ ,  $\{\alpha_i\}$ ,  $\{\beta_i\}$ ,  $\{\delta_i\}$ ,  $\{\gamma_{ij}\}$ ,  $\{\lambda_{ij}\}$  to zero.

Such models are invariant under the action of the relevant permutation group  $G = S_r \times S_2$  acting on the cells of the table by permutation of row/column levels and/or switching the row and column variable (McCullagh, 2000).

Potential models include well-known models such as symmetry, quasi-symmetry and quasi-independence.

A further invariant component can be obtained by replacing  $\{\alpha_i, \delta_i\}$  by  $\{\alpha_i \sin \phi, \alpha_i \cos \phi\}$  for any  $\phi \in (-\pi/2, \pi/2)$ .

For example,  $\phi = \tan^{-1}(1/2)$  allows the diagonal parameter model

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

(as well as the independence model)

Table 4: Distribution of marriages in the village of Nemgéné by lineage of each spouse (Cazes, 1990)

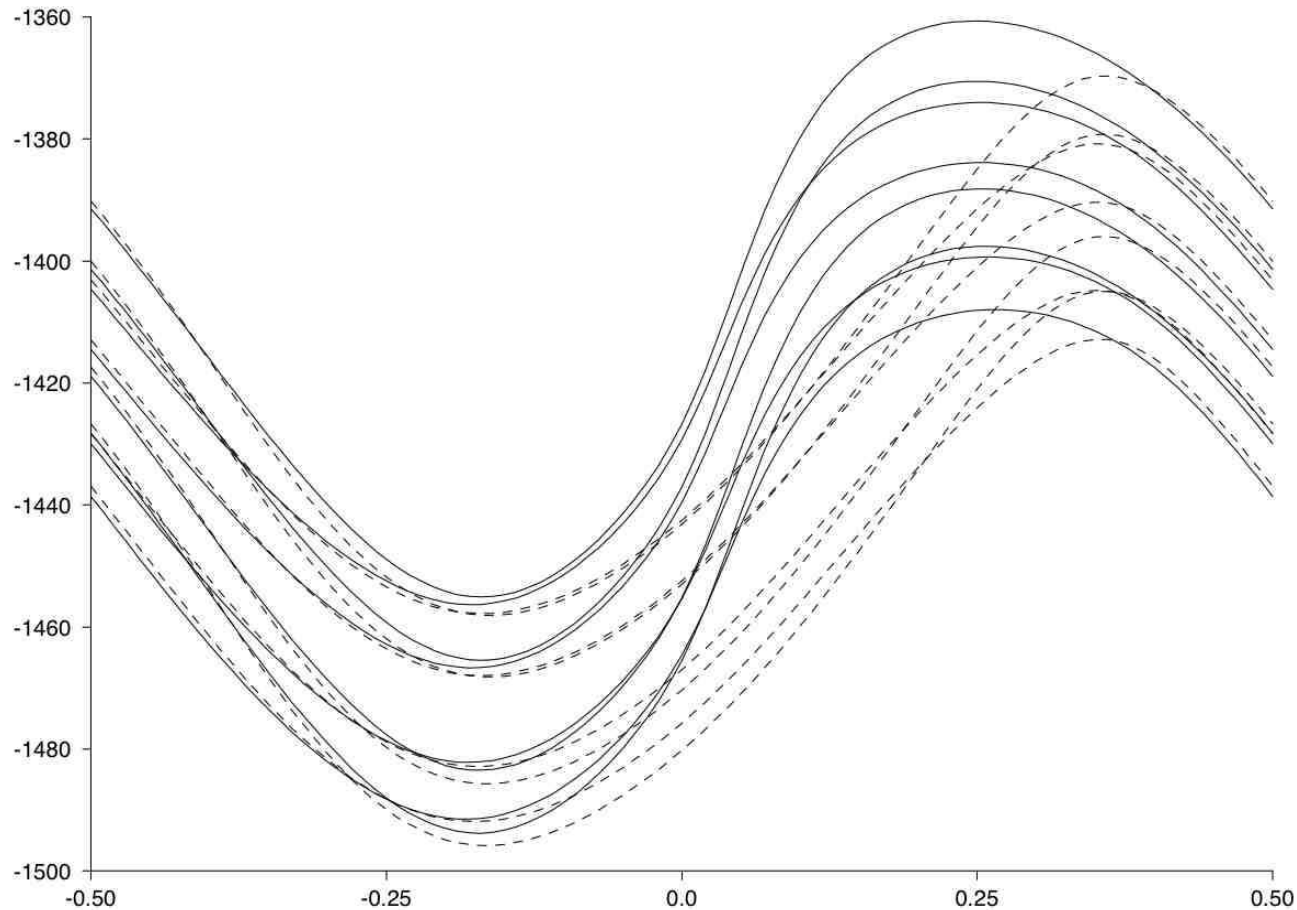
Husband's lineage	Wife's lineage								
	Iariwa Ger.	Segiwa	Suraba	Pussuwoï	Iariwa	Tengo <sub>2</sub>	Tengo	Other	
Iariwa Ger.	<b>0</b>	2	3	3	3	4	1	3	
Segiwa	2	<b>0</b>	2	5	1	1	6	3	
Suraba	0	2	<b>0</b>	2	3	6	7	6	
Pussuwoï	1	4	4	<b>0</b>	3	5	9	7	
Iariwa	2	2	3	4	<b>1</b>	1	13	16	
Tengo <sub>2</sub>	4	0	4	8	4	<b>1</b>	16	14	
Tengo	4	6	8	10	8	14	<b>12</b>	16	
Other	2	6	7	4	17	6	21	<b>17</b>	

Weak evidence for the models 'QI+Marginal Homogeneity' (posterior probability = 0.0132).

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

The preferred model (posterior probability = 0.9867) reflects some common structure of the diagonal and off-diagonal cells.

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



Log marginal likelihoods for each model with  $\phi$  plotted against  $\phi/\pi$ .  
Solid/dashed lines include/exclude  $\nu$ . In order of modal height:  
 $\emptyset, \{\beta\}, \{\gamma\}, \{\beta, \gamma\}, \{\lambda\}, \{\beta, \lambda\}, \{\gamma, \lambda\}, \{\beta, \gamma, \lambda\}$