

Bayesian Methods for Multivariate Categorical Data

Jon Forster
(University of Southampton, UK)

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

Table 2: Risk factors for coronary heart disease (Edwards and Havránek, 1985)

A 2^6 table (not displayed)

1841 men cross-classified by six risk factors for coronary heart disease

A: smoking

B: strenuous mental work

C: strenuous physical work

D: systolic blood pressure

E: ratio of α and β lipoproteins

F: family anamnesis of coronary heart disease

Table 3: Disclosure risk estimation - Large and sparse

Six potential key variables from the 3% Individual SAR for the 2001 UK Census (<http://www.ccsr.ac.uk/sars/2001>).

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.

The data

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

Individuals $i = 1, \dots, n$ are classified by variables $j = 1, \dots, p$, where variable j has k_j (potentially ordered) categories.

Independent categorical response vectors $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ip})$ are observed.

The contingency table \mathbf{y} , derived from $\{\mathbf{y}_i, i = 1, \dots, n\}$ has $k = \prod_1^p k_i$ cells.

Multinomial models

\mathbf{y} has a multinomial(n, π) distribution.

[or \mathbf{y} is drawn as a n/N sample from a finite population \mathbf{Y} which has a multinomial(N, π) *prior* distribution, Ericson, 1969]

$$\pi \in S_{p-1} = \left\{ \pi_j > 0, j = 1, \dots, p : \sum_{j=1}^p \pi_j = 1 \right\}$$

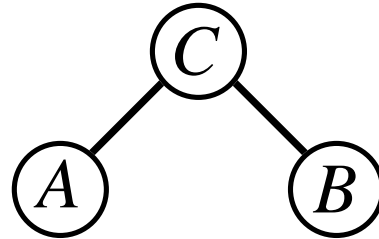
typically constrained by a model

e.g. assume a graphical or log-linear model for π ,

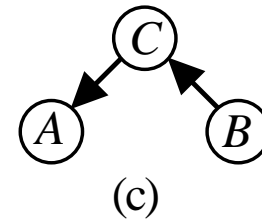
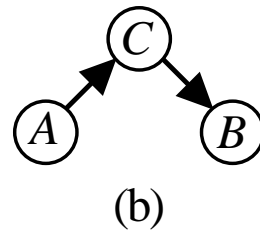
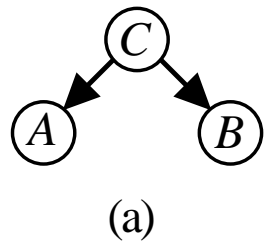
$$\log \pi = \mathbf{X}_m \boldsymbol{\beta}_m$$

Modelling association

Undirected graphical models



Directed graphical models



General log-linear models

Bayesian inference

The posterior for the model parameters β is obtained using

$$p(\beta|\mathbf{y}) \propto p(\mathbf{y}|\beta)p(\beta)$$

Posterior \propto likelihood \times prior

This is Bayes theorem.

Bayesian inference for the unconstrained model

For the unconstrained model $\beta = \pi$ and a natural prior for π is the Dirichlet with hyperparameters $\alpha = (\alpha_1, \dots, \alpha_p)$.

$$E(\pi_j) = \frac{\alpha_j}{\sum_{\ell=1}^p \alpha_\ell} \quad \text{Var}(\pi_j) = \frac{E(\pi_j)[1 - E(\pi_j)]}{1 + \sum_{\ell=1}^p \alpha_\ell}$$

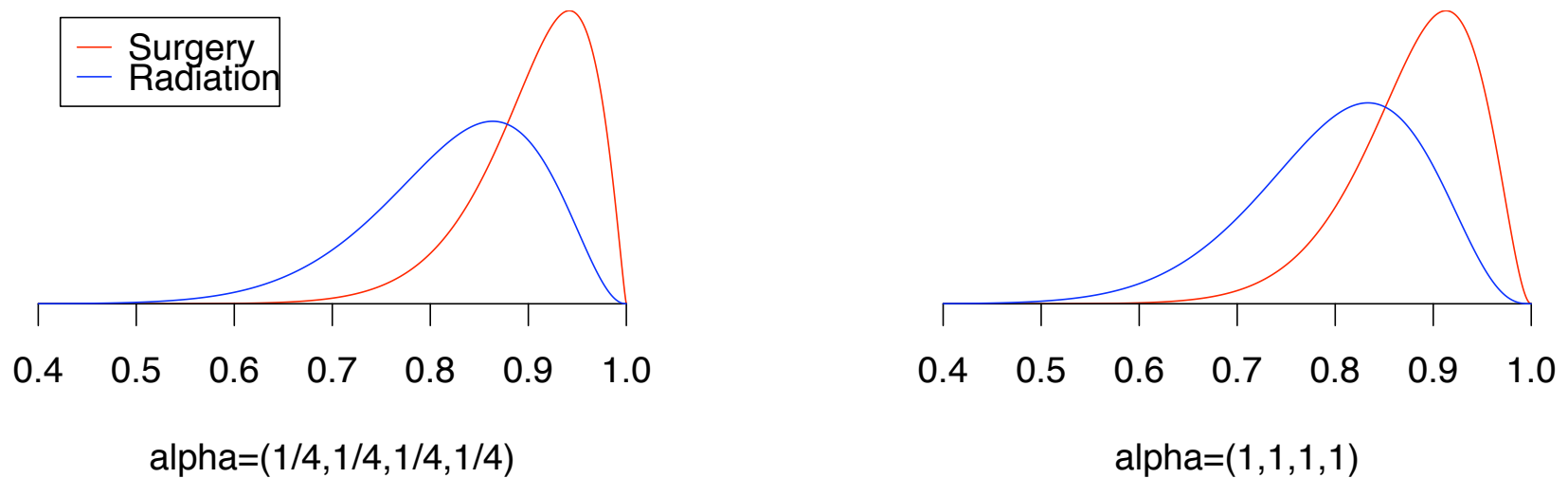
This prior distribution is *conjugate*, leading to a Dirichlet posterior with

$$\alpha \rightarrow \alpha + \mathbf{y}.$$

Possible diffuse priors have $\alpha = (1, \dots, 1)$, $\alpha = (\frac{1}{2}, \dots, \frac{1}{2})$, $\alpha = (\frac{1}{p}, \dots, \frac{1}{p})$.

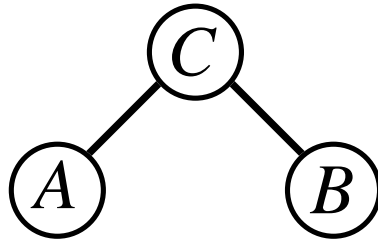
Treatment regimes for cancer of the larynx (from Agresti, 1990)

	Cancer controlled after treatment?	
	Yes	No
Surgery	21	2
Radiation Therapy	15	3



Posterior distributions for $P(\text{Yes}|\text{Surgery})$, $P(\text{Yes}|\text{Radiation})$.

Decomposable graphical models and the hyper-Dirichlet



A is independent of B given C , so that

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

or

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

Independent Dirichlet priors for $\{\beta_{i|k}^A\}$, $\{\beta_{j|k}^B\}$, for each k , and for $\{\beta_k^C\}$.

Conjugate prior leads to tractable computation.

Bayesian inference under model uncertainty

Allows model uncertainty to be coherently incorporated.

Multiple models indexed by $m \in M$.

Joint prior uncertainty about (m, β_m) is encapsulated by

$$p(m, \beta_m) = p(m)p(\beta_m|m)$$

where $p(m)$ is a discrete prior distribution over M .

By Bayes theorem

$$p(\beta_m|\mathbf{y}, m) = \frac{p(\mathbf{y}|m, \beta_m)p(\beta_m|m)}{p(\mathbf{y}|m)}$$

and

...

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

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

Note that, for any two models, say $m = 1$ and $m = 2$

$$\frac{p(m = 1|\mathbf{y})}{p(m = 2|\mathbf{y})} = \frac{p(m = 1)}{p(m = 2)} \frac{p(\mathbf{y}|m = 1)}{p(\mathbf{y}|m = 2)}$$

posterior odds = prior odds \times Bayes factor

The *Bayes factor* quantifies how belief about the two models is moderated in light of the observed data.

Model averaging

Model search and uncertainty is incorporated through the discrete prior [posterior] distribution $p(m)$ [$p(m|\mathbf{y})$] over the models $m \in M$.

Then, posterior predictive expectations of any function of \mathbf{Y} will be a *model average*

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

The posterior model probabilities may not be of interest in themselves – interpret them as weights.

Model uncertainty for laryngeal cancer data

	Cancer controlled after treatment?	
	Yes	No
Surgery	21	2
Radiation Therapy	15	3

$m = 1$: Independence of Outcome and Treatment; $\pi_{jk} = \beta_j^T \beta_k^O$

$m = 2$: Dependence of Outcome on Treatment; π unconstrained

Prior when $m = 1$: Independent Dirichlets on marginal probabilities, β_j^T, β_k^O

Prior when $m = 2$: Dirichlet on π

$$\text{Bayes factor (for independence)} = \begin{cases} 2.11 & \boldsymbol{\alpha} = (1, 1, 1, 1) \\ 3.17 & \boldsymbol{\alpha} = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) \\ 5.33 & \boldsymbol{\alpha} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \end{cases}$$

Computation for more complex data

Potential computational difficulties

1. Evaluating integrals – may be mathematically intractable
2. Number of models is large.

For decomposable models and hyper-Dirichlet prior distributions, most of the calculations (*e.g.* Bayes factors) can be performed exactly, and easily.

Otherwise MCMC or approximation such as BIC will be required.

For more than a few (3 or 4) cross-classifying variables, number of models gets large and some kind of approximation will be required. (*e.g.* Monte Carlo – Madigan and York, 1996)

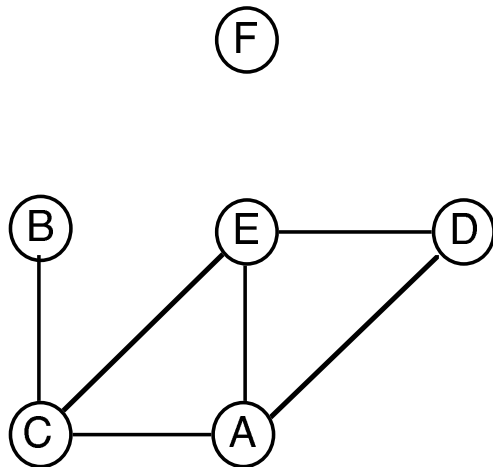
We adopt a search strategy to identify a *set of most probable models*

Table 1: Analysis

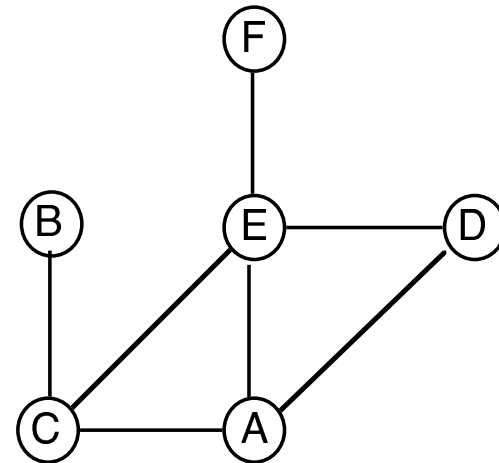
	Models			
	HO+HA	O+HA	A+HO	H+O+A
Hyper-Dirichlet (Jeffreys)	0.0360	0.0170	0.6428	0.3042
Hyper-Dirichlet (Perks)	0.0000	0.0001	0.0290	0.9709

Table 2: Analysis

A: smoking; B: strenuous mental work; C: strenuous physical work;
D: systolic blood pressure; E: ratio of α and β lipoproteins;
F: family anamnesis of coronary heart disease



(a) 0.3565 (0.0081)



(b) 0.1032 (0.0042)

Table 3: Analysis

Six potential key variables from the 3% Individual SAR for the 2001 UK Census (<http://www.ccsr.ac.uk/sars/2001>).

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.

Most probable model, based on sample data, is $SO+OAg+AgF+FN+NAc$

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	—	114	110	118	104	313	322	462	1543
	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

Record-level measures of disclosure risk

If E_j represents disclosure event in (sample non-empty) cell j

$$P(E_j|\mathbf{Y}) = \frac{1}{Y_j}$$

(Benedetti and Franconi, 1998)

Alternatively,

$$P(Y_j = 1|\mathbf{Y}) = I[Y_j = 1]$$

is the probability of uniqueness.

Bayesian disclosure risk assessment

We calculate Bayesian predictive probabilities as the posterior expectations

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

Hence our risk measures become

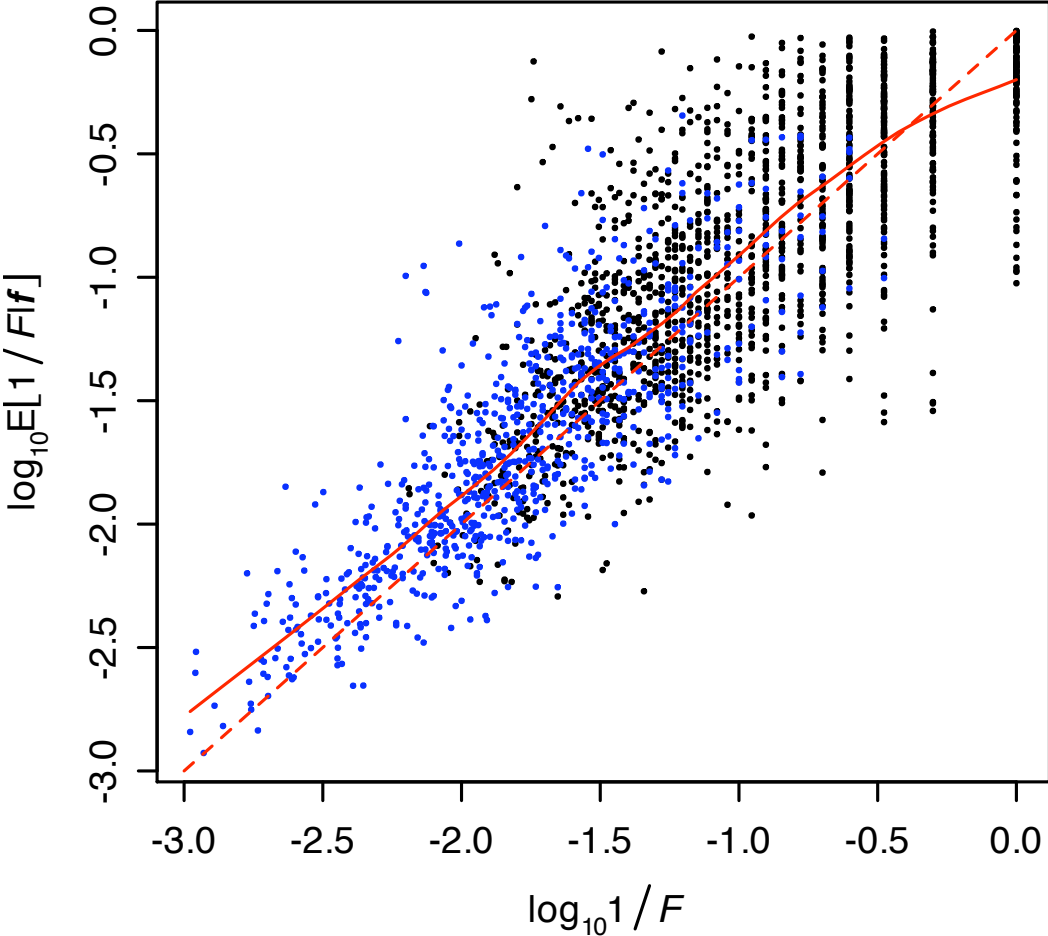
$$P(E_j|\mathbf{y}) = E[1/Y_j|\mathbf{y}]$$

and

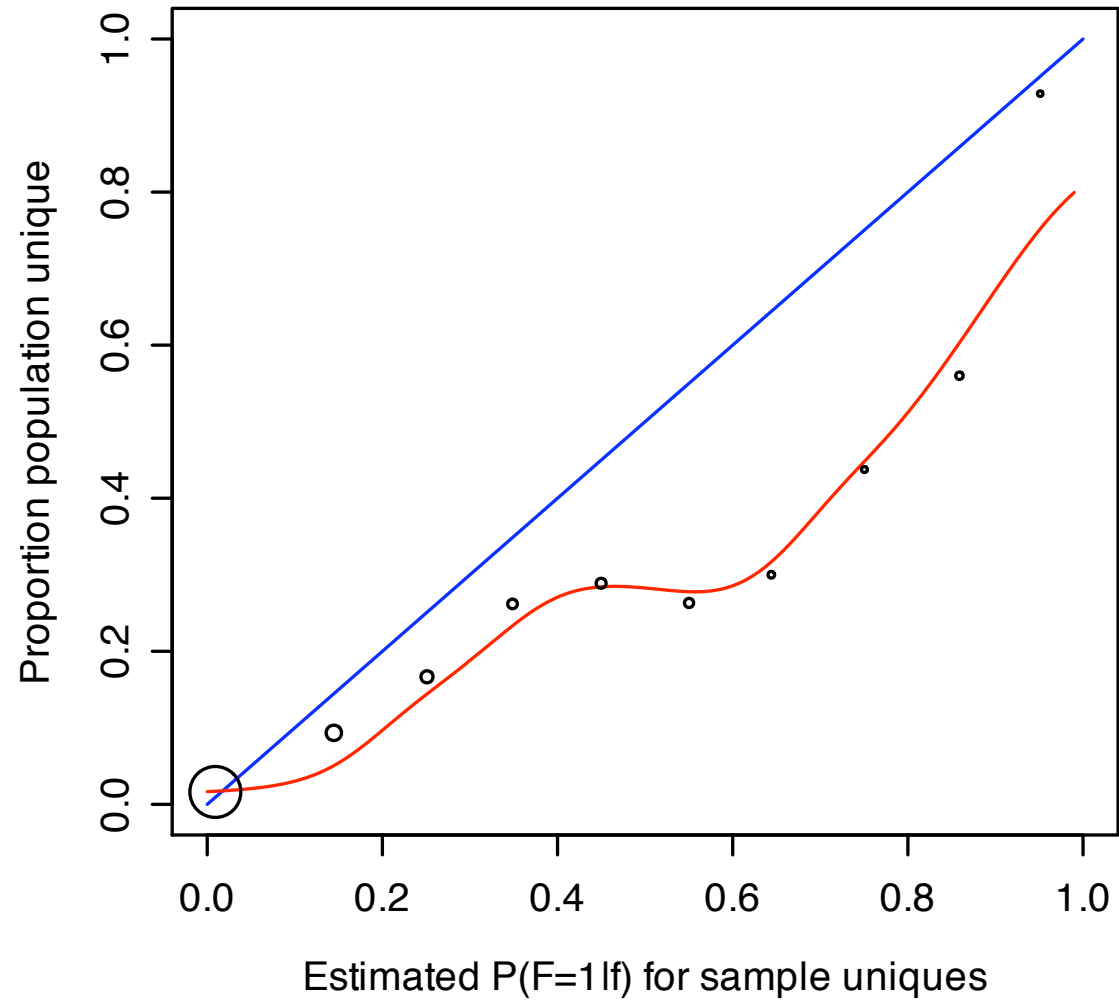
$$P(Y_j = 1|\mathbf{y}).$$

which are calculated using $P(\mathbf{Y} = \mathbf{y}|\mathbf{y})$.

Estimated v. True Disclosure Risk



True v. Estimated Probability of Uniqueness



ROC curve for uniqueness detection

