Some Thoughts on Prior Distributions and Posterior Model Probabilities

Ioannis Ntzoufras ,
Department of Business Administration, University of the Aegean, 8 Michalon Street, Island of Chios , Greece, e-mail: ntzoufras@aegean.gr.
Petros Dellaportas
Department of Statistics, Athens University of Economics and Business, 76 Patission Street, 10434, Athens, Greece, e-mail:petros@aueb.gr.
Jonathan J. Forster
Faculty of Mathematics, University of Southampton, UK, e-mail:jjf@maths.soton.ac.uk

Contents

1 Knuiman and Speed Example

- Knuiman and Speed (1988, Biometrics) Dataset
- $3 \times 2 \times 4$ Contingency Table
- 491 individuals classified by 3 categorical variables:
  - obesity (O: low,average,high)
  - hypertension (H:yes,no) and
  - alcohol consumption (A: 1,1–2,3–5,6+ drinks per day)
- Consider Poisson log-linear models to examine the association between them.

- Knuiman and Speed (1988), are setting rules for constructing meaningful prior distributions for the parameters of Poisson log-linear models used for inference in cross-tabulated data.
- Dellaportas and Forster (1999, Biometrika) have also used this dataset to illustrate Bayesian model selection using MCMC.
- Here, we incorporate the prior information of Knuiman and Speed (1988) in the model selection procedure.
- We illustrate results using a variety of prior distributions and adjusting dimensionality according to our desired penalty specification.

- The full Poisson log-linear model is given by

$$\begin{aligned} y_{ijk} &\sim Poisson(\lambda_{ijk}) \\ \log(\lambda_{ijk}) &= \beta_0 + \beta^O_i + \beta^H_j + \beta^A_k + \beta^{OH}_{ij} + \beta^{OA}_{ik} + \beta^{HA}_{jk} + \beta^{OHA}_{ijk} \end{aligned}$$

  for $i = 1, 2, 3$, $j = 1, 2$ and $k = 1, 2, 3, 4$ using sum-to-zero constraints.
- We use the general prior setup

$$\boldsymbol{\beta}_j \sim N\left(\boldsymbol{\mu}_j, \quad c^2_j \left(\boldsymbol{X}^T_j \boldsymbol{X}_j\right)^{-1}\right) \qquad (1)$$

- Initially we use two prior setups:
  1. Knuiman and Speed (1988) 'Informative Setup' used for Inference
  2. Dellaportas and Forster (1999) 'Low information' prior used for Bayesian Model Selection

1. Knuiman and Speed (1988) Prior:
   - Initial information
     - β^{OHA}_{ijk} and β^{OA}_{ik} are zero
     - β^{HA}_{jk} is non-zero with a priori estimated effects $\bar{\beta}^T_{HA} = (\beta^{HA}_{22}, \beta^{HA}_{23}, \beta^{HA}_{24}) = (-0.204, 0.088, 0.271)$.
   - Knuiman and Speed used a prior of type (1) with
     - $\boldsymbol{\mu}_{HA} = (\beta^{HA}_{22}, \beta^{HA}_{23}, \beta^{HA}_{24}) = (-0.204, 0.088, 0.271)$ and
     - $\boldsymbol{\mu}_j = \mathbf{0}$ for all $j \in \mathcal{V} \setminus \{HA\}$
     - $c^2_{OA} = c^2_{OHA} = 0$,
     - $c^2_{HA} = 0.05$ and
     - $c^2_j = \infty$ for $j \in \{\emptyset, O, H, A, OH\}$.
   - In order to avoid intractabilities in posterior model probabilities we adopt a slightly modified prior distribution with
     - $c^2_{OA} = c^2_{OHA} = 10^{-4}$,
     - $c^2_{HA} = 0.05$,
     - $c^2_j = 10^4$ for $j \in \{\emptyset, O, H, A, OH\}$

2. Dellaportas and Forster(1999) Prior:
   If no prior information is available then
   - $\boldsymbol{\mu}_j = \mathbf{0}$
   - Considered various choices for c_j:
     $c_j^2 = d, 2d, 4d$
     (d is the number of cells of the contingency table).
     Here we consider the choice $c_j^2 = 2d$.

The Uniform distribution on model space was a priori adopted.

Results were extracted using reversible jump MCMC methodology.

		$f(m\|\boldsymbol{y})$ DF	$f(m\|\boldsymbol{y})$ KS	KS prior information
1	O+H+A	0.680	0.056	no info
2	OH+A	0.315	0.000	no info
3	OA+H		0.056	zero
4	O+HA	0.003	0.443	non zero
5	OH+OA		0.000	zero
6	OH+HA	0.002	0.001	non zero
7	OA+HA		0.443	zero
8	OH+OA+HA		0.001	zero
9	OHA		0.000	zero

Table 1: Reversible Jump Estimated Posterior Model Probabilities (100,000 Iterations, Additional 10,000 Burn-in); DF= Dellaportas and Forster (1999) Prior, KS= Knuiman and Speed Prior.

		$f(Term\|\boldsymbol{y})$ DF	$f(Term\|\boldsymbol{y})$ KS	KS prior information
1	OH	0.317	0.002	no info
2	OA	0.000	0.500	zero
3	HA	0.005	0.888	non zero
4	OHA	0.000	0.000	zero

Table 2: Reversible Jump Estimated Posterior Term Probabilities (100,000 Iterations, Additional 10,000 Burn-in); DF= Dellaportas and Forster (1999) Prior, KS= Knuiman and Speed Prior.

Some Comments on Results

- Using DF Prior,
  - data support independence model (post.prob.=0.68)
  - Some support on the posterior significance of OH term (post.prob.=0.32).
- Using KS prior
  - OH term is not supported in contradiction to DF results
  - OA term is a posteriori supported by 50% [we cannot decide for its significance]. This is in contradiction to prior information and posterior results using DF prior
  - HA term is highly supported as a priori indicated [prior might be too strong]
  - OHA term is not supported [is in agreement with prior information and DF posterior results].

2 Expressing Posterior Model Odds as Penalised Information Criteria

Use more general setup than (1) given by

$$\boldsymbol{\beta}_m \sim N\left(\boldsymbol{\mu}_m, \quad \boldsymbol{C}_m \boldsymbol{\Sigma}_m \boldsymbol{C}_m\right) \qquad (2)$$

where

- m: model indicator
- $\boldsymbol{C}_m = Diag(c_{m,j} \boldsymbol{I}_{d_{m,j}})$
- $c_{m,j}$ is a variance multiplicator controlling the prior information for model parameters
- $d_{m,j}$ is the dimension of j term in m model
- $\boldsymbol{\Sigma}_m$ is a base variance - covariance matrix

Then $log f(m|y) =$

$$= C + \log f(y|m, \tilde{\beta}_m) - \frac{1}{2}(\tilde{\beta}_m - \mu_m)^T \boldsymbol{C}_m^{-1} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{C}_m^{-1} (\tilde{\beta}_m - \mu_m) - \frac{1}{2}\psi_m$$

$$\psi_m = \sum_{j \in m} d_{m,j} \log c_{m,j}^2 + \log|\boldsymbol{\Sigma}_m| + \log|\boldsymbol{C}_m^{-1} \boldsymbol{\Sigma}_m^{-1} \boldsymbol{C}_m^{-1} - H(\tilde{\beta}_m)| - 2\log f(m).$$

* C: constant
* $\tilde{\beta}_m$ is the posterior mode
* $H(\beta_m)$: second derivative matrix for $\log f(y|m, \beta_m)$

Interesting cases ($f(m) \propto 1$):

- $\boldsymbol{\Sigma}_m = (-H(\beta_m))^{-1}$, $c_{m,j} = c_m$ then $\psi_m = d_m \log c_m^2$
  - $c_m^2 = n$: Unit information prior (BIC penalty)
- $\boldsymbol{\Sigma}_m = (-H(\beta_m))^{-1}$, $H(\beta_m)$ diagonal, $\psi_m = \sum_{j \in m} d_{m,j} \log(c_{m,j}^2 + 1)$

If we a priori penalise by F for each additional parameter added in the model then

$$f(m) \propto e^{-F d_m/2}$$

resulting to

$$\psi_m = \sum_{j \in m} d_{m,j}(\log c^2_{m,j} + F) + \log|\boldsymbol{\Sigma}_m| + \log|\boldsymbol{C}_m^{-1}\boldsymbol{\Sigma}_m^{-1}\boldsymbol{C}_m^{-1} - H(\tilde{\beta}_m)|.$$

If we desire to imply posterior penalty $\psi_m = \log p_m$ the prior model odds should be specified by

$$f(m) \propto \sqrt{p_m^{-1}|\boldsymbol{C}_m^T\boldsymbol{\Sigma}_m\boldsymbol{C}_m||\boldsymbol{C}_m^{-1}\boldsymbol{\Sigma}_m^{-1}\boldsymbol{C}_m^{-1} - H(\tilde{\beta}_m)|}. \qquad (3)$$

If the prior base matrix $\boldsymbol{\Sigma}_m$ is equal to the Fisher information matrix then

$$f(m) \propto \sqrt{p_m^{-1}|\boldsymbol{C}_m^T\boldsymbol{C}_m + \boldsymbol{I}|} = \sqrt{p_m^{-1}\prod_{j \in m}(c^2_{m,j}+1)^{d_{m,j}}}.$$

Using the above prior model probabilities results to

$$\psi_m = \log p_m \quad + \quad \sum_{j \in m} d_{m,j} \log\left(\frac{c^2_{m,j}}{c^2_{m,j}+1}\right)$$
$$+ \quad \log|\boldsymbol{\Sigma}_m| + \log|\boldsymbol{C}_m^{-1}\boldsymbol{\Sigma}_m^{-1}\boldsymbol{C}_m^{-1} - H(\tilde{\beta}_m)|.$$

Advantages:

- Bounded penalty function for $c_{m,j} \to \infty$ [avoid Lindley's paradox].

$$\psi_m \to \log p_m + \log|\boldsymbol{\Sigma}_m| + \log| - H(\tilde{\beta}_m)|.$$

- Use informative prior within each model
- The penalty function is expressed as sum of
  - prior parameter p_m and
  - a distance measure between prior base matrix and posterior variance covariance function.
- Prior base matrix may be specified to have determinant equal to posterior covariance matrix.

3 More Results on the Example

Use two new prior setups:

Use Knuiman and Speed Prior within each model and

$$f(m) \propto \sqrt{p_m^{-1}|\boldsymbol{C}_m^T\boldsymbol{C}_m + \boldsymbol{I}|} \qquad (4)$$

with

$$\log(p_m) = \sum_{j \in m} d_j F_j. \qquad (5)$$

1. $F_j = \log(2d)$ for all terms (following Dellaportas and Forster arguments)
2. $F_j = \log(2d)$ for $j \neq HA$ and $F_{HA} = \log(2)$ [small penalty equal to two data points].

		$f(m\|y)$			
		DF	KS	KS	KS
1	O+H+A	0.680	0.056	0.624	0.144
2	OH+A	0.315	0.000	0.298	0.070
3	OA+H		0.056		
4	O+HA	0.003	0.443	0.057	0.533
5	OH+OA		0.000		
6	OH+HA	0.002	0.001	0.024	0.253
7	OA+HA		0.443		0.000
8	OH+OA+HA		0.001		0.000
9	OHA		0.000		
	$f(m) \propto$	1	1	(4) & (5)	(4) & (5)
	$F_j,\ j \neq HA$	-	-	$\log(2d)$	$\log(2d)$
	F_{HA}	-	-	$\log(2d)$	$\log(2)$

Table 3: Reversible Jump Estimated Posterior Model Probabilities (100,000 Iterations, Additional 10,000 Burn-in).

		$f(Term\|\boldsymbol{y})$			
		DF	KS	KS	KS
1	OH	0.317	0.002	0.322	0.323
2	OA	0.000	0.500	0.000	0.000
3	HA	0.005	0.888	0.081	0.786
4	OHA	0.000	0.000	0.000	0.000
	$f(m) \propto$	1	1	(4) & (5)	(4) & (5)
	$F_j,\ j \neq HA$	-	-	$\log(2d)$	$\log(2d)$
	F_{HA}	-	-	$\log(2d)$	$\log(2)$

Table 4: Reversible Jump Estimated Posterior Term Probabilities (100,000 Iterations, Additional 10,000 Burn-in); DF= Dellaportas and Forster (1999) Prior, KS= Knuiman and Speed Prior.

Comments on Results

- Posterior model probabilities using KS prior and prior model probabilities defined by (4)& (5) are similar to Dellaportas and Forster results. Differences are due to prior information within each model.
- Prior information on the significance of a term may be expressed by using lower penalty without affecting the significance of the other terms.

4 Discussion

- The specification of Prior distributions is Important for Bayesian Model Selection
- Why not express our beliefs for models via prior penalties?
- Divide model selection procedure in:
  (a) Estimation (prior of $\boldsymbol{\beta}_{(m)}$)
  (b) Model selection (penalize to support parsimony principle).