Aspects of Bayesian Model Choice

Ioannis Ntzoufras[1], Petros Dellaportas [2]

Department of Statistics, Athens University of Economics and Business,

76 Patission Street, 10434, Athens, Greece

Jonathan J. Forster [3]

Faculty of Mathematics, University of Southampton, UK

May 1998

For postcript file:

http://stat-athens.aueb.gr/~jbn/ntzoufras.html#papers

[1]E-mail:jbn@stat-athens.aueb.gr

[2]E-mail:petros@aueb.gr

[3]E-mail:jjf@maths.soton.ac.uk

1 The Problem

- We investigate some aspects of a well known problem: The effect of model parameters prior distributions on posterior odds.
- The question is: how we can use the relationship between posterior odds and information criteria to eliminate the effect of prior distributions on posterior model probabilities.

In this presentation we investigate ...

- ... the effect of a "model independent" Normal prior with diagonal prior variance on a nested model comparison.
- ... the effect of a "model dependent" Normal prior on a more general model comparison.
- ... the association of the posterior odds with information criteria.
- ... how we can calibrate the prior model weights in order to achieve a desired penalty imposed on the log-likelihood.
- ... how similar Normal priors affect the Bayes factor estimated by SSVS introduced by George and McCullogh (1993) in a nested model comparison.

2 Information Criteria.

Generally, most information criteria select the model that minimize a quantity similar to

$$IC_m = -2log\left(f(\mathbf{y}|\hat{\theta}_m, m)\right) + d_m F \tag{1}$$

- θ_m is the parameter vector and $\hat{\theta}_m$ are the the MLE.
- F is the penalty for each additional parameter used in the model.

In linear regression models:

- $\theta_m^T = [\beta_{(m)}^T, \sigma^2]$.
- Minimizing $-2log\left(f(\mathbf{y}|\hat{\theta}_m, m)\right)$ is equivalent to minimizing $nlog(RSS_m)$; RSS_m is the residual sum of squares of model m.

Special cases of (1) are given for different F

- For $F = 2$ we have AIC, Akaike (1973).
- For $F = log(n)$ we have BIC, Schwarz (1978).
- For $F = clog(log(n))$ we have Φ_c by Hannan and Quinn (1979).
- For $F = 3/2$ we have the criterion of Smith and Spiegelhalter (1980).

In pairwise comparisons of models m_0 and m_1 we define as $IC_{01} = IC_0 - IC_1$ resulting in

$$IC_{01} = -2log\left(\frac{f(\mathbf{y}|\hat{\theta}_0, m_0)}{f(\mathbf{y}|\hat{\theta}_1, m_1)}\right) - (d_1 - d_0)F \qquad (2)$$

assuming that $d_0 < d_1$.

It is obvious that if

- $IC_{01} < 0$ we prefer the simpler model m_0
- $IC_{01} > 0$ we prefer the more complicated model m_1.

Also note that we have the same support pattern when we use $-2log(PO_{01})$

- $-2log(PO_{01}) < 0$ supports the simpler model m_0
- $-2log(PO_{01}) > 0$ supports the more complicated model m_1.

where PO_{01} are the posterior odds of model m_0 vs. model m_1.

3 Model Independent Normal Prior with Diagonal Covariance Matrix in Nested Models Comparison.

Let us have two regression models (m_0 and m_1) where $y_i \sim N(\mu_i, \sigma^2)$ with

$$\mathbf{H}_0 : \mu = \mathbf{z}\delta \quad vs. \quad \mathbf{H}_1 : \mu = \mathbf{z}\delta + \mathbf{x}\beta.$$

where

- $\mathbf{z}$ is a $n \times p$ data/design matrix containing information about $Z_1, \ldots, Z_k$ common terms; $\mathbf{z} = (\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_k})$,
- $\delta^T = (\delta_1^T, \delta_2^T, \ldots, \delta_k^T)$; δ_i model coefficient of Z_i term.
- $\mathbf{x}$ is a $n \times d$ data/design matrix of variable/term X,
- β is the vector of the model coefficients of term X
- $\mu^T = (\mu_1, \mu_2, \ldots, \mu_n)$ is a $n \times 1$ vector.
- $d_0 = p$ and $d_1 = p + d$.

Prior Distributions:

$$\delta|\sigma^2 \sim N(0, \mathbf{V}\sigma^2),\ \mathbf{V} = Diag(c^2(\mathbf{z_i^T z_i})^{-1}) \tag{3}$$

$$\beta|\sigma^2 \sim N(0, c^2(\mathbf{x^T x})^{-1}\sigma^2) \tag{4}$$

$$\sigma^{-2} \sim \Gamma(a_0, b_0)$$

Similar priors were proposed by Raftery *et al.* (1997) for generalised linear models.

For $a_0 = b_0 = 0$ we have

$$B_{01} = \frac{f(\mathbf{y}|m_0)}{f(\mathbf{y}|m_1)} = (c^2)^{d/2} |\mathbf{x}^\mathbf{T}\mathbf{x}|^{-1/2} |\mathbf{E}|^{-1/2} \left(\frac{SS_0}{SS_1}\right)^{-n/2} \quad (5)$$

where

- $\mathbf{V} = Diag(c^2(\mathbf{z_i^T z_i}))$, $\mathbf{A}^{-1} = \mathbf{z^T z} + \mathbf{V}^{-1}$,
- $\mathbf{\Delta}^{-1} = (\mathbf{I} - \mathbf{zAz^T})$, $\mathbf{E}^{-1} = \mathbf{x}^T \mathbf{\Delta}^{-1} \mathbf{x} + c^{-2} \mathbf{x^T x}$,
- $SS_0 = \mathbf{y}^T(\mathbf{I} - \mathbf{zAz^T})\mathbf{y} = \mathbf{y}^T \mathbf{\Delta}^{-1} \mathbf{y}$,
- $SS_1 = \mathbf{y}^T(\mathbf{\Delta}^{-1} - \mathbf{\Delta^{-1} x E x^T \Delta^{-1}})\mathbf{y}$.

Then the posterior odds can be written similarly to (2)

$$-2log(PO_{01}) = nlog\left(\frac{SS_0}{SS_1}\right) - dF$$

with

$$F = log(c^2) - \frac{1}{d} log|\mathbf{x^T x}| - \frac{1}{d} log|\mathbf{E}| - \frac{2}{d} log\left(\frac{p(m_1)}{p(m_0)}\right). \quad (6)$$

Usually we use large c and therefore

$$-2log\left(\frac{f(\mathbf{y}|\hat{\theta}_0, m_0)}{f(\mathbf{y}|\hat{\theta}_1, m_1)}\right) = nlog\left(\frac{RSS_0}{RSS_1}\right) \approx nlog\left(\frac{SS_0}{SS_1}\right)$$

resulting in

$$IC_{01} \approx -2log(PO_{01})$$

with F given by (6) and

$$IC_{01} \approx -2log(B_{01})$$

with F given by (6) and $p(m_0) = p(m_1)$.

SPECIAL CASE: $\mathbf{x}, \mathbf{z}$ orthogonal and $p(m_0) = p(m_1)$ then $F = log(c^2 + 1)$.

In this model comparison we note that

- The prior parameter c which controlls the prior variance of model parameters has large influence on Bayes factor.
- Large values of c imply a flat prior for model parameters, but then the Bayes factor strongly supports the simpler model!
- The penalty is also affected the determinants $|\mathbf{x}^T\mathbf{x}|$ and $|\mathbf{E}|$.
- We can have some control of the imposed penalty via the prior parameter c.
- The prior parameter c and the prior model odds have different influence of the penalty F.

$$\begin{array}{llllllll} p(m_1)/p(m_0) & \uparrow & \text{F} & \downarrow & PO_{01} & \downarrow \\ c^2 & \uparrow & \text{F} & \uparrow & PO_{01} & \uparrow \end{array}$$

3.1 A Simple Example

Let us take the simple possible case where $p = d = 1$. The resulting penalty (6) using the $p(m_0) = p(m_1)$ is equal to

$$F = log(c^2+1) + log\left(1 - \left(\frac{c^2}{c^2+1}\right)^2 r^*_{xz}\right)$$

with

$$r^*_{xz} = \frac{\left(r_{xz} + \frac{\bar{x}}{\sigma_x}\frac{\bar{z}}{\sigma_z}\right)^2}{(1+\frac{\bar{x}^2}{\sigma_x^2})(1+\frac{\bar{z}^2}{\sigma_z^2})} = \frac{\left(r_{xz} + \frac{\bar{x}}{\sigma_x}\frac{\bar{z}}{\sigma_z}\right)^2}{\left(1 + \frac{\bar{x}}{\sigma_x}\frac{\bar{z}}{\sigma_z}\right)^2 + \left(\frac{\bar{x}}{\sigma_x} - \frac{\bar{z}}{\sigma_z}\right)^2} \leq 1$$

where r_{xz} corellation coeficients of X and Z with $r_{xz} \leq 1$.
Consider the extreme case where $x = \lambda z$ then

- $r_{xz} = 1$, that is both variables carry exactly the same information and only one of them should be included in the model.
- Also $r^*_{xz} = 1$ resulting in $F = log\left(\frac{2c^2+1}{c^2+1}\right)$.
- For large c, $F \approx log2$ resulting in Bayes factor of 1.41 (58.5% over 41.5%) in favour of the simpler model whatever is the relation between z and y!
- Possible ways to avoid the effect of collinearity on posterior odds is to use the orthogonal transformed data, see Clyde *et al.* (1996), Clyde (1998), or use the model dependent priors used by Smith and Kohn (1996) and George and Foster (1997).

4 Model Dependent Normal Prior.

Let us have two regression models (m_0 and m_1) where $y_i \sim N(\mu_i, \sigma^2)$ with

$$\mathbf{H}_0 : \mu = \mathbf{x}_0\beta_{(0)} \; vs. \; \mathbf{H}_1 : \mu = \mathbf{x}_1\beta_{(1)}.$$

where

- $\mathbf{x}_m$ is the design/data matrix of model m, with dimesion $n \times d_m$
- $\beta_{(m)}$ is the $d_m \times 1$ parameter vector of model m.

Prior Distributions:

$$\beta_{(m)}|\sigma^2 \sim N(0, c^2(\mathbf{x}_m^T\mathbf{x}_m)^{-1}\sigma^2) \tag{7}$$

$$\sigma^{-2} \sim \Gamma(a_0, b_0)$$

These priors were used by Smith and Kohn (1996) and George and Foster (1997).

Using improper prior for σ^2 we have

$$B_{01} = \frac{f(\mathbf{y}|m_0)}{f(\mathbf{y}|m_1)} = (c^2+1)^{(d_1-d_0)/2}\left(\frac{SS_0}{SS_1}\right)^{-n/2}$$

As in the previous section we have

$$-2log(PO_{01}) = nlog\left(\frac{SS_0}{SS_1}\right) - (d_1-d_0)F$$

with

$$F = log(c^2+1) - \frac{2}{d_1-d_0}log\left(\frac{p(m_1)}{p(m_0)}\right). \tag{8}$$

For large c we have again

$$IC_{01} \approx -log(PO_{01})$$

when F is given by (8).

Therefore this prior

- results in Bayes factor that is not affected by collinearity of explanatory variables.
- allows us to fully control the penalty of the log-likelihood via c.
- c and the prior odds have different influence on the penalty (and the posterior odds).

Note that this prior cannot be computed if an expanatory variable is an exact linear function of the others. In such extreme case we may remove this term from the model.

5 Calibration of Log-likelihood Penalty through Prior Probabilities

The idea is:

- The 'naive' uniform prior on model space should be avoided. George and Foster (1997) noted that this prior in variable selection models supports models with high dimensionality.
- The prior parameter c and the prior odds have different influence on the penalty F (and posterior odds) and therefore their combination may not express our real prior belief for models. An idea is that the prior odds should be defined as a function of c, or vice versa.
- Here we suggest to define prior probabilities in such way that we eliminate the effect of prior parameter c and impose a desired penalty. This idea was used by Poskitt and Treymane (1983) for Jefreys prior. Similar approach was used by Pericchi (1984) using normal priors.

For the model comparison of section 4 we may use

$$p(m) \propto \left(\frac{c^2+1}{\alpha} \right)^{d_m/2}$$

so (8) results in penalty $F = log(\alpha)$. The parameter α now fully controlls the penalty that will be imposed on the log-likelihood. Note that we can easily adopt this kind of prior to variable selection problems using binary indicator variables γ as in George and McCullogh (1993) SSVS. In this case when there are no restrictions on the model space we use

$$\frac{p(\gamma_i = 1)}{p(\gamma_i = 0)} = \left(\frac{c^2+1}{\alpha} \right)^{p_i/2}$$

where p_i are the number of parameters for the i term.

For the model comparison of section 3 we may use

$$\frac{p(m_1)}{p(m_0)} = \left(\frac{c^2}{\alpha} \right)^{d/2} |\mathbf{E}|^{-1/2} |\mathbf{x}^T \mathbf{x}|^{-1/2}$$

resulting again in penalty $F = log(\alpha)$.

This prior can be easily adopted in pairwise comparisons but it seems very difficult to use it in MCMC model selection methods. The effect of c can be eliminated even in MCMC methods but priors that a-priori assume independence of model parameters will be affected by associations of the explanatory variables.

6 Comparison of Nested Models with SSVS.

6.1 Model Independent Normal Prior with Diagonal Covariance Matrix

George and McCullogh (1993) proposed to use a Gibbs sampler for variable selection where the non-significant parameter is not set to zero but has a distribution near zero.

We consider the model selection of section 3 with as similar priors as possible. Therefore we use (3) for δ and

$$\beta \sim \gamma N(0, c^2(\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}\sigma^2) + (1-\gamma)N(0, \kappa^{-2}c^2(\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}\sigma^2) \qquad (9)$$

where $\gamma = 1$ when β is in the model (model m_1) and $\gamma = 0$ otherwise (model m_0).

$$-2log(B_{01}^{SSVS}) = nlog\left(\frac{SS_0^{SSVS}}{SS_1}\right) - dF^{SSVS}$$

$$F^{SSVS} = log(c^2) + log\left(\frac{\kappa^2}{\kappa^2-1}\right) - \frac{1}{d}log\left(\frac{|\mathbf{x}^T\mathbf{x} + \frac{c^2}{\kappa^2-1}\mathbf{E}^{-1}|}{|\mathbf{E}^{-1}|}\right)$$

For $c^2 < \kappa^2 \to \infty$, $F^{SSVS} \to F$ and $SS_0^{SSVS} \to SS_0$. Therefore $B_{01}^{SSVS} \to B_{01}$ which is also noted by George and McCullogh (1997). Prior odds may be appropriately selected to have $F^{SSVS} = log(\alpha)$.

6.2 Model Dependent Normal Prior.

We try to use model dependent priors with SSVS. We apply the following prior on model comparison of section 3

$$\begin{bmatrix} \delta \\ \beta \end{bmatrix} \sim N(0, c^2 \mathbf{D}_0 (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{D}_0 \sigma^2)$$

with

$$\mathbf{G} = [\mathbf{z},\ \mathbf{x}]$$

and

$$\mathbf{D}_0 = \begin{bmatrix} \mathbf{I}_p & 0 \\ 0 & \kappa^{\gamma-1}\mathbf{I}_d \end{bmatrix}$$

where $\mathbf{I}_p$ is the identity matrix with dimension p; $\gamma = 0, 1$. The prior of model m_1 is the same as (7), resulting in

$$-2log(B_{01}^{SSVS}) = nlog\left(\frac{SS_0^{SSVS}}{SS_1^{SSVS}}\right) - dF^{SSVS}$$

$$F^{SSVS} = log(c^2+1) + log\left(\frac{\kappa^2}{c^2+\kappa^2}\right)$$

$$-\frac{1}{d}log\left(\frac{|\mathbf{x}^T\mathbf{x} - \frac{(c^2+\kappa)^2}{(c^2+\kappa^2)(c^2+1)}\mathbf{x}^T\mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{x}|}{|\mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{z}(\mathbf{z}^T\mathbf{z})^{-1}\mathbf{z}^T\mathbf{x}|}\right)$$

Note that when $c^2 < \kappa^2 \to \infty$ then $SS_\gamma^{SSVS} \to SS_\gamma$ and $F^{SSVS} \to F = log(c^2+1)$. Therefore, for large κ^2 we similar results as in section 4. Note that Ntzoufras *et al.* (1998) for log-linear models suggest $\kappa = 1000$.

7 Conclusions.

- The uniform prior on model space can be avoided.
- The combination of the prior odds and the prior variance of model parameters may not reflect our real prior beliefs for models and may strongly support more complicated or simpler model than a-priori desired.
- Smith and Kohn (1996) prior seems to be a good choice for linear models leading to simple interpreted posterior odds. Prior probabilities should be determined as function of dimensionality and of a new parameter which controls the imposed penalty.
- SSVS for suitable priors approximates the usual Bayes factor. The effect of prior variance may also eliminated in SSVS using appropriate prior weights.
- Generalization can be made for GLM through Laplace approximations.

References

[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*, 267–281, Budapest: Akademia Kiado.

[2] Clyde, M.A. (1998). Bayesian Model Averaging and Model Search Strategies *Bayesian Statistics* 6, to appear.

[3] Clyde, M.A., DeSimone, H. and Parmigiani, G. (1996). Prediction via Orthognalized Model Mixing. *Journal of the American Statistical Association*, **91**, 1197–1208.

[4] George, E.I. and Foster, D.P. (1997). Calibration and Empirical Bayes Variable Selection. *Technical Report*, University of Pensilvania, USA.

[5] George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.

[6] George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339–373.

[7] Hannan, E.J. and Quinn, B.G. (1979). The Determination of the Order of an Autoregression . *Journal of the Royal Statistical Society*, B, **41**, 190–195.

[8] Ntzoufras, I., Forster, J.J. and Dellaportas, P. (1998). Stochastic Search Variable Selection for Log-linear Models. *Technical Report* 38, Department of Statistics, Athens University of Economics and Business, Greece.

[9] Pericchi, L.R. (1984). An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models. *Biometrika*, **71**, 575–586.

[10] Poskitt, D.S. and Tremayne, A.R. (1983). On the Posterior Odds of Time Series Models. *Biometrika*, **70**, 157–162.

[11] Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, **92**, 179–191.

[12] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.

[13] Smith, A.F.M. and Spiegelhalter, D.J.(1980). Bayes Factor and Choice Criteria for the Linear Models. *Journal of the Royal Statistical Society*, B, **42**, 213–220.

[14] Smith M. and Kohn R. (1996). Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics*, **75**, 317–343.