SPECIFICATION AND INTERPRETATION OF PRIOR DISTRIBUTIONS FOR VARIABLE SELECTION IN LINEAR MODELS

Ioannis Ntzoufras*, Petros Dellaportas* and Jonathan J. Forster†

[e-mail for correspondence: jbn@stat-athens.aueb.gr]

Abstract

This paper investigates the implication of flat priors in model selection and their influence on posterior odds. The association of posterior odds and information criteria is also examined. Some easily calibrated prior setups are proposed in order to use flat prior distributions on model parameters and impose the desired a-priori penalty to the log-likelihood.

Keywords: GLM, Bayesian Model Selection, Priors, Variable Selection.

1 Introduction

This work examines 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 a comment of Lindley's paradox Bartlett (1957) noted that for any fixed dataset if we increase the prior variance of the parameters then the Bayes factor of model m_0 *vs.* model m_1, B_{01}, fully supports the null model. The investigation of this case is of prime interest since traditionally used flat priors cannot be used.

The traditional model selection tool is the log-likelihood ratio used under several strategies (backward, forward, stepwise). Likelihood ratio test is problematic since for large sample sizes selects highly complicated models. In 1970's Akaike (1973), Schwarz (1978), Hannan and Quinn (1979) and many others introduced the use minimum criteria based on log-likelihood test correcting for bias by penalising for extra parameters used. The aim of such criteria was the selection of similar models when the sample size increases. These criteria differ in the balance they give between parsimony and goodness of fit.

On the other hand, Bayesian model selection is based either on posterior odds and Bayes factors (mainly when model selection is viewed as hypothesis testing), see Kass and Raftery (1995) for an excellent review, or

*Department of Statistics, Athens University of Economics and Business, Greece.

†Faculty of Mathematics, University of Southampton, UK.

in maximising expected utility; see Bernardo and Smith (1994). Both cases may result to similar information criteria penalising the log-likelihood ratio in different ways depending on prior distribution or utility used; see Schwarz (1978), Smith and Spiegelhalter (1980), Poskitt and Treymane (1983), Poskitt (1987), Kass and Wasserman (1995), Pauler (1998), Key *et al.* (1998) and Bernardo (1998).

In this paper we consider the posterior odds approach and investigate some priors that may be used in linear models in order to achieve the desired penalty. We claim that a 'non-informative' prior distribution of model parameters should reflect prior beliefs only for model parameters and should not affect the model selection procedure. Moreover, prior probabilities should be the tool to express prior belief on model selection.

2 Information Criteria.

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

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

where $\boldsymbol{\theta}_m$ is the parameter vector and $\hat{\boldsymbol{\theta}}_m$ are the the maximum likelihood estimates, $d(m)$ is the dimension of the model m and F is the penalty for each additional parameter used in the model. In linear regression models we have that $\boldsymbol{\theta}_m^T = [\boldsymbol{\beta}_{(m)}^T, \sigma^2]$ and minimizing $-2log\left(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_m, m)\right)$ 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 equal to 2, $log(n)$, $c\ log(log(n))$ and 3/2 we have AIC by Akaike (1973), BIC by Schwarz (1978), Φ_c by Hannan and Quinn (1979) and criterion of Smith and Spiegelhalter (1980). There are many other criteria with a variety of penalties but we will limit ourselves in the most broadly used.

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

$$IC_{01} = -2log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_0, m_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_1, m_1)}\right) - (d(m_1) - d(m_0))F \tag{2}$$

assuming that $d(m_0) < d(m_1)$.

It is obvious that if $IC_{01} < 0$ we prefer the simpler model m_0 and when $IC_{01} > 0$ we prefer the more complicated model m_1. Also note that we have the same support pattern when we use $-2log(PO_{01})$ and therefore $-2log(PO_{01}) < 0$ supports the simpler model m_0 while $-2log(PO_{01}) > 0$ supports the more complicated model m_1; where PO_{01} are the posterior odds of model m_0 vs. model m_1. Moreover, there is a variety of publications connecting specific information criteria with posterior odds and Bayes factor; see Schwarz (1978), Smith and Spiegelhalter (1980), Poskitt and Tremayne (1983), Kass and Wasserman (1995), Pauler (1998) and Key *et al.* (1998). Most of the information criteria where mainly introduced for model selection in linear models and therefore a modification may be needed for the rest GLMs. Pauler (1998) proposes some modifications for BIC based on similar arguments of Schwarz (1978) for the linear models.

Generally we can write the information criteria in more general setup such as

$$IC_{01} = -2log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_0, m_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_1, m_1)}\right) - \psi, \tag{3}$$

where ψ is a penalty function depending on difference of model dimensionalities, $d(m_1) - d(m_0)$, sample size n, and design matrices, $\boldsymbol{x}_{(m_0)}$ and $\boldsymbol{x}_{(m_1)}$.

3 Priors for Normal Linear Models

In this section we investigate the resuting posterior odds in linear models using various Normal prior distributions for model parameters. Both conjugate and non-conjugate Normal prior distributions are used and their relation to information criteria in noted. The linear model is given by

$$\boldsymbol{y} \sim N_n(\boldsymbol{x}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n) \tag{4}$$

where $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and dimensionality k, n is number of observations, $\boldsymbol{y}$ is a $n \times 1$ vector of response data, $\boldsymbol{x}_{(m)}$ and $\boldsymbol{\beta}_{(m)}$ are the the $n \times d(m)$ data matrix and the $d(m) \times 1$ parameter vector of model m, σ^2 is the error variance of the linear model, $\boldsymbol{I}_k$ is the unit matrix with dimension k.

In this section we use conjugate Normal prior for model parameters. We use the following hierarchical prior for the model coefficients. This prior is convinient for the calculation of the posterior odds since we can intergrate out all model parameters. we may use prior

$$\boldsymbol{\beta}_{(m)}|\sigma^2 \sim N_{d(m)}(\boldsymbol{\mu}_{(m)}, \boldsymbol{V}_{(m)}\sigma^2) \tag{5}$$

where $\boldsymbol{\mu}_{(m)}$ is the prior mean of the parameters of model m, and $\boldsymbol{V}_{(m)}$ is a matrix controling the prior covariance matrix of model parameters. The prior on σ^{-2} can be the usual gamma with parameters a_0 and b_0. Here we use the improper prior with $a_0 = b_0 = 0$. We call this prior conjugate because the resulting conditional posterior distributions $f(\boldsymbol{\beta}_{(m)}|\sigma^{-2}, m, \boldsymbol{y})$ and $f(\sigma^{-2}|\boldsymbol{\beta}_{(m)}, m, \boldsymbol{y})$ are the same family as their coresponding priors.

The resulting posterior odds for comparing model m_0 and model m_1 is given by

$$PO_{01} = \frac{p(m_0)}{p(m_1)} \left(\frac{|V^{-1}_{(m_0)}|}{|V^{-1}_{(m_1)}|} \right)^{1/2} \left(\frac{|\boldsymbol{E}_{(m_1)}|}{|\boldsymbol{E}_{(m_0)}|} \right)^{1/2} \left(\frac{SS_1}{SS_0} \right)^{n/2} \tag{6}$$

where

$$SS_m = \boldsymbol{y}^T\boldsymbol{y} - \tilde{\boldsymbol{\beta}}^T_{(m)}\boldsymbol{E}_{(m)}\tilde{\boldsymbol{\beta}}_{(m)} + \boldsymbol{\mu}^T_{(m)}\boldsymbol{V}^{-1}_{(m)}\boldsymbol{\mu}_{(m)}$$

$$\tilde{\boldsymbol{\beta}}_{(m)} = \boldsymbol{E}^{-1}_{(m)}(\boldsymbol{x}^T_{(m)}\boldsymbol{x}_{(m)}\hat{\boldsymbol{\beta}}_{(m)} + \boldsymbol{V}^{-1}_{(m)}\boldsymbol{\mu}_{(m)}),\ \boldsymbol{E}_{(m)} = \boldsymbol{x}^T_{(m)}\boldsymbol{x}_{(m)} + \boldsymbol{V}^{-1}_{(m)}$$

An alternative notation is given by Atkinson (1978) and Pericchi (1984) where

$$SS_m = RSS_m + (\hat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\mu}_{(m)})^T((\boldsymbol{x}^T_{(m)}\boldsymbol{x}_{(m)})^{-1} + \boldsymbol{V}_{(m)})^{-1}(\hat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\mu}_{(m)})$$

where $RSS_m = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T_{(m)}(\boldsymbol{x}^T_{(m)}\boldsymbol{x}_{(m)})\hat{\boldsymbol{\beta}}_{(m)}$ is the residual sum of squares. The above quantity the sum of two measures: goodness of fit and the difference between maximum likelihood estimates and the prior distribution.

3.1 Special case 1: Independent Priors for Each Term Parameter Vector.

We usually prefer to use a simpler prior covariance matrix. Moreover, in generalised linear models it is convinient to handle the problem as variable selection and use similar setup as Dellaportas *et al.* (1998). The data matrix and parameter vector of the full model can be divided is submatrices $\boldsymbol{x}_i$ and subvectors $\boldsymbol{\beta}_i$ corespoding to i factor, variable or interaction term. A convinient set up is to use independent normal priors on these subvectors. Therefore use the prior $\boldsymbol{\beta}_i|\sigma^2 \sim N(0, c^2(\boldsymbol{x}_i^T\boldsymbol{x}_i)^{-1}\sigma^2)$. Raftery *et al.* (1997) proposeded similar prior for GLM. This prior results to the following posterior odds

$$PO_{01} = \frac{p(m_0)}{p(m_1)}(c^2)^{(d(m_1)-d(m_0))/2}\prod_{i=1}^{p}|\boldsymbol{x}_i^T\boldsymbol{x}_i|^{(\gamma_{0,i}-\gamma_{1,i})/2}\left(\frac{|\boldsymbol{E}_{(m_1)}|}{|\boldsymbol{E}_{(m_0)}|}\right)^{1/2}\left(\frac{SS_1}{SS_0}\right)^{n/2}$$

where

$$SS_m = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}_{(m)}\boldsymbol{x}_{(m)}^T\boldsymbol{x}_{(m)}\boldsymbol{E}_{(m)}^{-1}\boldsymbol{x}_{(m)}^T\boldsymbol{x}_{(m)}\hat{\boldsymbol{\beta}}_{(m)}$$

where p is the total number of terms of the full model, $\boldsymbol{\gamma}_m$ is the usual vector of variable indicators for model m used in variable selection problems by George and McCullogh (1993), Foster and George (1994), George and Foster (1997), Dellaportas *et al.* (1998), Clyde (1998) and Kuo and Mallick (1998), and $\gamma_{m,i}$ is the value of the indicator of i term for model m.

The matrix $\boldsymbol{E}_{(m)}$ is given by

$$\boldsymbol{E}_{(m)} = \boldsymbol{D}_{\gamma_m}^T\boldsymbol{E}\boldsymbol{D}_{\gamma_m}$$

where $\boldsymbol{D}_{\gamma_{m,i}}$ is the $d(f)\times d(m)$ matrix created by the $Diag(\gamma_i\boldsymbol{I}_{d_i})$ matrix if we eliminate all columns of with $\gamma_{m,i}=0$, where $d(f)$ is the dimension of the full model, d_i is the number of parameters that correspond to the i term and $\boldsymbol{E}$ corresponds to the full model with

$$\boldsymbol{E} = \begin{bmatrix} \frac{c^2+1}{c^2}\boldsymbol{x}_1^T\boldsymbol{x}_1 & \dots & \boldsymbol{x}_1^T\boldsymbol{x}_i & \dots & \boldsymbol{x}_1^T\boldsymbol{x}_p \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_i^T\boldsymbol{x}_1 & \dots & \frac{c^2+1}{c^2}\boldsymbol{x}_i^T\boldsymbol{x}_i & \dots & \boldsymbol{x}_i^T\boldsymbol{x}_p \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_p^T\boldsymbol{x}_1 & \dots & \boldsymbol{x}_p^T\boldsymbol{x}_i & \dots & \frac{c^2+1}{c^2}\boldsymbol{x}_p^T\boldsymbol{x}_p \end{bmatrix}$$

It is obvious that for c^2 large then the quantity $(c^2+1)/c^2$ involved in the diagonal of $\boldsymbol{E}$ becomes one and therefore $\boldsymbol{E}_{(m)} \approx \boldsymbol{x}_{(m)}^T\boldsymbol{x}_{(m)}$ for all models. In this case

$$SS_m \approx \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}_{(m)}(\boldsymbol{x}_{(m)}^T\boldsymbol{x}_{(m)})\hat{\boldsymbol{\beta}}_{(m)} = RSS_m$$

Therefore, for this prior setup and with c^2 large the $-2log(PO_{01})$ can be approximately written as (3) with associated penalty function

$$\psi = (d(m_1)-d(m_0))log(c^2) + \sum_{i=1}^{p}(\gamma_{0,i}-\gamma_{1,i})log|\boldsymbol{x}_i^T\boldsymbol{x}_i| + log\left(\frac{|\boldsymbol{E}_{(m_1)}|}{|\boldsymbol{E}_{(m_0)}|}\right) + 2log\left(\frac{p(m_0)}{p(m_1)}\right). \tag{7}$$

Note that in the special case where the design/data matrix is orthogonal then this penalty simplifies to $\psi = (d(m_1)-d(m_0))log(c^2+1) + 2log\left(\frac{p(m_0)}{p(m_1)}\right)$.

Now let us consider the usual case of comparison two nested models m_0 and m_1 that differ only in the i term (that is $\gamma_{m_0,j}=\gamma_{m_1,j}$ for all $j\neq i$ and $\gamma_{m_0,i}=0$, $\gamma_{m_1,i}=1$). Then the resulted penalty is given by

$$\psi = d_i log(c^2+1) - log|\boldsymbol{x}_i^T\boldsymbol{x}_i| + log|\boldsymbol{x}_i^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{x}_i| + 2log\left(\frac{p(m_0)}{p(m_1)}\right) \tag{8}$$

$$\boldsymbol{\Delta}_{(m_0)} = \boldsymbol{I}_n - \frac{c^2}{c^2+1}\boldsymbol{x}_{(m_0)}\boldsymbol{E}_{(m_0)}^{-1}\boldsymbol{x}_{(m_0)}^T$$

The matrix $\boldsymbol{x}_i^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{x}_i$ involved in the determinant, for $c^2\to\infty$ is approximately equal to the residual sum of square of a regression with response the additional variable i and explanatory all the variables of the simpler model m_0.

3.2 Special case 2: Model Dependent Normal Prior.

In this section we will use as the prior used by Smith and Kohn (1996) and George and Foster (1997) for variable selection where

$$\beta_{(m)}|\sigma^2, m \sim N(0, c^2(\boldsymbol{x}_{(m)}^T \boldsymbol{x}_{(m)})^{-1}\sigma^2) \tag{9}$$

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

Using improper prior for σ^2 we have

$$PO_{01} = \frac{p(m_0)}{p(m_1)}\frac{f(\boldsymbol{y}|m_0)}{f(\boldsymbol{y}|m_1)} = (c^2+1)^{(d(m_1)-d(m_0))/2}\left(\frac{SS_0}{SS_1}\right)^{-n/2}$$

Resulting to equation similar to (3) with penalty

$$\psi = (d(m_1) - d(m_0))log(c^2+1) - 2log\left(\frac{p(m_1)}{p(m_0)}\right). \tag{10}$$

For large c^2 we have again $IC_{01} \approx -log(PO_{01})$ when ψ is given by (10). Therefore, this prior results to Bayes factor that is not affected by collinearity of explanatory variables, allows to fully control the penalty of the log-likelihood via c^2 Both c^2 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. Consider now the case of comparison of two models that differ only in one explanatory variable involving a single parameter. In the case where the additional variable is highly correlated with the other variables in the model m_0 (e.g. regresion $R^2 = 0.99$ of the new variable over the others) then the likelihood ratio test will be very small and therefore the Bayes factor will approximately be c^2+1 in favour of the simpler model. Therefore, the prior parameter c^2 in this prior setup has two straightforward interpretations. The first is closely related to the the penalty attributed to the log-likelihood ratio for each additional parameter used and the second is related to the Bayes factor of the model m_0 vs. the same model with one additional variable which is highly correlated with the rest of the variables in the model.

4 Discussion

This paper presents some problems and possible solutions when we use flat non-informative priors in Bayesian model selection. Some specific model selection setups in linear and generalized linear models were presented. The connection of posterior odds and information criteria is also reported.

Generally, we can argue that 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. Therefore, a joint calibration of both prior odds and prior variance is needed. 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.

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] Atkinson, A.C. (1978). Posterior Probabilities for Choosing a Regression Model. *Biometrika*, **65**, 39–48.

[3] Bartlett, M.S. (1957). Comment on 'A Statistical Paradox' by D.V. Lindley. *Biometrika*, **44**, 533–534.

[4] Bernardo, J.M. (1998). Nested Hypothesis Testing: the Bayesian Reference Criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[5] Bernardo, J.M. and Smith, A.F.M. (1994) Bayesian Theory. Chichester: John Wiley & Sons.

[6] Clyde, M.A. (1998). Bayesian Model Averaging and Model Search Strategies. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[7] Dellaportas, P., Forster, J.J. and Ntzoufras, I.(1998). On Bayesian Model and Variable Selection Using MCMC. *Technical Report* **40**, Department of Statistics, Athens University of Economics and Business, Greece.

[8] Foster, D.P. and George, E.I. (1994). The Risk Inflation Criterion for Multiple Regression. *Annals of Statistics*, **22**, 1947–1975.

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

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

[11] 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.

[12] Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

[13] Kass, R.E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

[14] Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1998). Bayesian Model Choice: What and Why? *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[15] Kuo, L. and Mallick, B. (1998). Variable Selection for Regression models. *Sankhyā A*, , (to appear).

[16] Pauler, D.K. (1998). The Schwarz Criterion and Related Methods for Normal Linear Models. *Biometrika*, **85**, 13–27.

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

[18] Poskitt, D.S. (1987). On the Posterior Odds of Time Series Models. *Journal of the Royal Statistical Society B*, **49**, 199–208.

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

[20] 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.

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

[22] 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.

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