Specification of prior distributions under model uncertainty

Petros Dellaportas[1], Jonathan J. Forster[2] and Ioannis Ntzoufras[3]

September 3, 2008

SUMMARY

We consider the specification of prior distributions for Bayesian model comparison, focusing on regression-type models. We propose a particular joint specification of the prior distribution across models so that sensitivity of posterior model probabilities to the dispersion of prior distributions for the parameters of individual models (Lindley's paradox) is diminished. We illustrate the behavior of inferential and predictive posterior quantities in linear and log-linear regressions under our proposed prior densities with a series of simulated and real data examples.

Keywords: Bayesian inference; BIC; Generalised linear models; Lindley's Paradox; Model averaging; Regression models;

1 Introduction and motivation

A Bayesian approach to inference under model uncertainty proceeds as follows. Suppose that the data $\boldsymbol{y}$ are considered to have been generated by a model m, one of a set M of competing models. Each model specifies the distribution of $\boldsymbol{Y}$, $f(\boldsymbol{y}|m,\boldsymbol{\beta}_m)$ apart from an unknown parameter vector $\boldsymbol{\beta}_m \in B_m$, where B_m is the set of all possible values for the coefficients of model m. We assume that $B_m = \mathcal{R}^{d_m}$ where d_m is the dimensionality of $\boldsymbol{\beta}_m$.

If $f(m)$ is the prior probability of model m, then the posterior probability is given by

$$f(m|\boldsymbol{y}) \;=\; \frac{f(m)f(\boldsymbol{y}|m)}{\sum\limits_{m\in M} f(m)f(\boldsymbol{y}|m)}, \qquad m \in M \tag{1}$$

where $f(\boldsymbol{y}|m)$ is the marginal likelihood calculated using $f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|m,\boldsymbol{\beta}_m)f(\boldsymbol{\beta}_m|m)d\boldsymbol{\beta}_m$ and $f(\boldsymbol{\beta}_m|m)$ is the conditional prior distribution of $\boldsymbol{\beta}_m$, the model parameters for model m. Therefore

$$f(m|\boldsymbol{y}) \;\propto\; f(m)f(\boldsymbol{y}|m), \qquad m \in M.$$

For any two models m_1 and m_2, the ratio of the posterior model probabilities (posterior odds in favour of m_1) is given by

$$\frac{f(m_1|\boldsymbol{y})}{f(m_2|\boldsymbol{y})} \;=\; \frac{f(m_1)}{f(m_2)}\frac{f(\boldsymbol{y}|m_1)}{f(\boldsymbol{y}|m_2)} \tag{2}$$

[1] Department of Statistics, Athens University of Economics and Business, Greece ,email:petros@aueb.gr

[2] Department of Mathematics, University of Southampton, email:jjf@maths.soton.ac.uk

[3] Department of Statistics, Athens University of Economics and Business, Greece, email:ntzoufras@aueb.gr

the ratio of prior probabilities multiplied by the ratio of marginal likelihoods, also known as the Bayes factor.

The posterior distribution for the parameters of a particular model is given by the familiar expression

$$f(\boldsymbol{\beta}_m|m,\boldsymbol{y}) \;\propto\; f(\boldsymbol{\beta}_m|m)f(\boldsymbol{y}|\boldsymbol{\beta}_m,m), \qquad m\in M.$$

For a single model, a highly diffuse prior on the model parameters is often used (perhaps to represent ignorance). Then the posterior density takes the shape of the likelihood and is insensitive to the exact value of the prior density function, provided that the prior is relatively flat over the range of parameter values with non-negligible likelihood. When multiple models are being considered, however, the use of such a prior may create an apparent difficulty. The most obvious manifestation of this occurs when we are considering two models m_1 and m_2 where m_1 is completely specified (no unknown parameters) and m_2 has parameter $\boldsymbol{\beta}_{m_2}$ and associated prior density $f(\boldsymbol{\beta}_{m_2}|m_2)$. Then, *for any observed data* $\boldsymbol{y}$, the Bayes factor in favour of m_1 can be made arbitrarily large by choosing a sufficiently diffuse prior distribution for $\boldsymbol{\beta}_{m_2}$ (corresponding to a prior density $f(\boldsymbol{\beta}_{m_2}|m_2)$ which is sufficiently small over the range of values of $\boldsymbol{\beta}_{m_2}$ with non-negligible likelihood). Hence, under model uncertainty, two different diffuse prior distributions for model parameters might lead to essentially the same posterior distributions for those parameters, but very different Bayes factors.

This result was discussed by Lindley (1957) and is often referred to as 'Lindley's paradox' although it is also variously attributed to Bartlett (1957) and Jeffreys (1961). As Dawid (1999) points out, the Bayes factor is only one of the two elements on the right hand side of (2) which contribute towards the posterior model probabilities. The prior model probabilities are of equal significance. By focusing on the impact of the prior distributions for model parameters on the Bayes factor, there is an implicit understanding that the prior model probabilities are specified independently of these prior distributions. This is often the case in practice, where a uniform prior distribution over models is commonly adopted, as a reference position. Examples where non-uniform prior distributions have been suggested include Madigan et al (1995), Chipman (1996), Laud and Ibrahim (1995, 1996), and Chipman et al (2001). In this paper, we consider how the two elements of the prior distribution under model uncertainty might be jointly specified so that perceived problems with Bayesian model comparison can be avoided.

A related issue concerns the use of improper prior distributions for model parameters. Such prior distributions involve unspecified constants of proportionality, which do not appear in posterior distributions for model parameters but do appear in the marginal likelihood for any model and in any associated Bayes factors, so these quantities are not uniquely determined. There have been several attempts to address this issue, and to define an appropriate Bayes factor for comparing models with improper priors; see Kadane and Lazar (2004) for a review. In such examples, Dawid (1999) proposes that the product of the prior model 'probability' and the prior density for a given model could be determined simultaneously by eliciting the relative prior 'probabilities' of particular sets of parameter values for different models. He also suggests an approach for constructing a general non informative prior, over both models and model parameters, based on Jeffreys' priors for individual models. Although the prior distributions for individual models are not generally proper, they have densities which are uniquely determined and hence the posterior distribution over models can be evaluated. Here, we do not consider improper prior distributions for the model parameters, but our approach is similar in spirit as we do explicitly consider a joint specification of the prior over models and model parameters.

We focus on models in which the parameters are sufficiently homogeneous (perhaps after transformation) that a multivariate normal prior density $N(\boldsymbol{\mu}_m, V_m)$ is appropriate, and in which the likelihood is sufficiently regular for standard asymptotic results to apply. Examples are linear regression models, generalized linear models and standard time series models. In much of what follows, with minor modification, the normal prior can be replaced by any elliptically symmetric prior density proportional to $|V|^{-1/2} g\left((\boldsymbol{\beta}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)$ where $\int_0^\infty r^{d-1} g(r^2) dr < \infty$ and d is the dimensionality of $\boldsymbol{\beta}$. This includes prior distributions from the multivariate t or Laplace families.

We choose to decompose the prior variance matrix as $V_m = c_m^2 \Sigma_m$ where c_m represents the scale of the prior dispersion and Σ_m is a matrix with a specified value of $|\Sigma_m|$; for example $|\Sigma_m| = 1$, although in what follows we will not use an explicit value. Hence, suppose that

$$f(\boldsymbol{\beta}_m|m) = (2\pi)^{-d_m/2} |\Sigma_m|^{-1/2} c_m^{-d_m} \exp\left(-\frac{1}{2c_m^2}(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)\right). \tag{3}$$

Then,

$$\begin{aligned} f(m|\boldsymbol{y}) &\propto f(m) \int f(\boldsymbol{y}|m, \boldsymbol{\beta}_m) f(\boldsymbol{\beta}_m|m) d\boldsymbol{\beta}_m \\ &= f(m)(2\pi)^{-d_m/2} |\Sigma_m|^{-1/2} c_m^{-d_m} \times \\ &\quad \int_{\mathcal{R}^{d_m}} \exp\left(-\frac{1}{2c_m^2}(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)\right) f(\boldsymbol{y}|m, \boldsymbol{\beta}_m) d\boldsymbol{\beta}_m \end{aligned} \tag{4}$$

and for suitably large c_m,

$$f(m|\boldsymbol{y}) \approx f(m)(2\pi)^{-d_m/2} |\Sigma_m|^{-1/2} c_m^{-d_m} \int_{\mathcal{R}^{d_m}} f(\boldsymbol{y}|m, \boldsymbol{\beta}_m) d\boldsymbol{\beta}_m. \tag{5}$$

Hence, as c_m^2 gets larger, $f(m|\boldsymbol{y})$ gets smaller, assuming everything else remains fixed. Therefore, for two models of different dimension with the same value of c_m^2, the posterior odds in favor of the more complex model tends to zero as c_m^2 gets larger, that is as the prior dispersion increases at a common rate. This is essentially Lindley's paradox.

There have been substantial recent computational advances in methodology for exploring the model space, see for example Green (1995, 2003), Kohn et al (2001), Denison et al (2002), Hans et al (2007). The related discussion of the important problem of choosing prior parameter dispersions has been largely focused on ways to avoid Lindley's paradox; see, for example, Fernandez et al (2001) and Liang et al (2008) for detailed discussion on appropriate choices of Zellner's g-priors for linear regression models and Raftery (1996) and Dellaportas and Forster (1999) for some guidelines on selecting dispersion parameters of normal priors for generalized linear model parameters. The important effect that these prior specifications might have on the parameter posterior distributions within each model has been neglected. For example, a set of values of c_m might be appropriate for addressing model uncertainty, but might produce prior densities $f(\boldsymbol{\beta}_m|m)$ that are insufficiently diffuse and overstate prior information within certain models. This has a serious effect on posterior and predictive densities of all quantities of interest in any data analysis.

In this paper we propose that prior distributions for model parameters should be specified with the issue of inference conditional on a particular model being the primary focus. For example, when only weak information concerning the model parameters is available, a highly diffuse prior may be deemed appropriate. The key element of our proposed approach is that sensitivity of posterior model probabilities to the exact scale of such a diffuse prior is avoided by suitable specification of prior model probabilities $f(m)$. As mentioned above, these probabilities are rarely specified carefully, a discrete uniform prior distribution across models usually being adopted. However, it is straightforward to see that setting $f(m) \propto c_m^{d_m}$ in (5) will have the effect of eliminating dependence of the posterior model probability $f(m|y)$ on the prior dispersion c_m. This provides a motivation for investigating how prior model probabilities can be chosen in conjunction with prior distributions for model parameters, by first considering properties of the resulting posterior distribution.

2 Prior and posterior distributions

We consider the joint specification of the two components of the prior distribution by investigating its impact on the asymptotic posterior model probabilities. By using Laplace's method to approximate the posterior marginal likelihood in (4), we obtain, subject to certain regularity conditions (see, Kass et al, 1988, Schervish, 1995, sec. 7.4.3),

$$\begin{aligned} f(m|\boldsymbol{y}) \propto f(m)|\Sigma_m|^{-1/2} c_m^{-d_m} f(\boldsymbol{y}|m,\widehat{\boldsymbol{\beta}}_m) \exp\left(-\frac{1}{2c_m^2}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\mu}_m)^T\Sigma_m^{-1}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\mu}_m)\right) \times \\ |c_m^{-2}\Sigma_m^{-1} - H(\widehat{\boldsymbol{\beta}}_m)|^{-1/2}\left(1+O_p(n^{-1})\right) \end{aligned} \qquad (6)$$

where $\widehat{\boldsymbol{\beta}}_m$ is the maximum likelihood estimate and $H(\boldsymbol{\beta}_m)$ is the second derivative matrix for $\log f(\boldsymbol{y}|m,\boldsymbol{\beta}_m)$. Then,

$$\begin{aligned} \log f(m|\boldsymbol{y}) &= C + \log f(m) - \frac{1}{2}\log|\Sigma_m| - d_m \log c_m + \log f(\boldsymbol{y}|m,\widehat{\boldsymbol{\beta}}_m) \\ &\quad -\frac{1}{2c_m^2}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\mu}_m)^T\Sigma_m^{-1}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\mu}_m) - \frac{1}{2}\log|c_m^{-2}\Sigma_m^{-1} - H(\widehat{\boldsymbol{\beta}}_m)| + O_p(n^{-1}) \\ &= C + \log f(m) - \frac{1}{2}\log|\Sigma_m| - d_m \log c_m + \log f(\boldsymbol{y}|m,\widehat{\boldsymbol{\beta}}_m) \\ &\quad -\frac{1}{2c_m^2}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\mu}_m)^T\Sigma_m^{-1}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\mu}_m) - \frac{d_m}{2}\log n - \frac{1}{2}\log|i(\widehat{\boldsymbol{\beta}}_m)| + O_p(n^{-1/2}) \end{aligned} \qquad (7)$$

where C is a normalizing constant to ensure that the posterior model probabilities sum to one and $i(\boldsymbol{\beta}_m) \approx -n^{-1}H(\boldsymbol{\beta}_m)$ is the Fisher information matrix for a unit observation; see Kass and

Wasserman (1995). If the decomposition of the prior variance matrix $c_m^2 \Sigma_m$ is chosen so that $|\Sigma_m| = |i(\boldsymbol{\beta}_m)|^{-1}$, then

$$\begin{aligned} \log f(m|\boldsymbol{y}) \quad = \quad & C + \log f(\boldsymbol{y}|m, \widehat{\boldsymbol{\beta}}_m) - \frac{1}{2c_m^2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m) \\ & + \log f(m) - d_m \log c_m - \frac{d_m}{2} \log n + O_p(n^{-1/2}) \end{aligned} \tag{8}$$

and c_m^{-2} can be interpreted as the number of units of information in the prior, defined as

$$c_m^{-2} = (|V_m||i(\boldsymbol{\beta}_m)|)^{-1/d_m} \,. \tag{9}$$

Note that substituting $c_m = 1$ (unit information) into (8), and choosing a discrete uniform prior distribution across models, suggests model comparison on the basis of a modified version of the Schwarz criterion (BIC; Schwarz, 1978) where maximum likelihood is replaced by maximum penalized likelihood. In a comparison of two nested models, Kass and Wasserman (1995) give extra conditions on a unit information prior which lead to model comparison asymptotically based on BIC; see Volinsky and Raftery (2000) for an example of the use of unit information priors for Bayesian model comparison. For regression-type models where the components of $\boldsymbol{y}$ are not identically distributed, depending on explanatory data, the unit information as defined above potentially changes as the sample size changes, so a little care is required with asymptotic arguments. We assume that the explanatory variables arise in such a way that $i(\boldsymbol{\beta}_m) = i_{\lim}(\boldsymbol{\beta}_m) + O(n^{-1/2})$ where $i_{\lim}(\boldsymbol{\beta}_m)$ is a finite limit. This is not a great restriction and is true, for example, where the explanatory data may be thought as i.i.d. observations from a distribution with finite variance.

In general, $i(\boldsymbol{\beta}_m)$ depends on the unknown model parameters, so the number of units of information c_m^{-2} corresponding to any given prior variance matrix V_m, will also not be known, and hence it is not generally possible to construct an exact unit information prior. Dellaportas and Forster (1999) and Ntzoufras et al (2003) advocated substituting $\boldsymbol{\mu}_m$, the prior mean of $\boldsymbol{\beta}_m$ into $i(\boldsymbol{\beta}_m)$ to give a prior for model comparison which has a unit information interpretation but for which model comparison is not asymptotically based on BIC.

When the prior distribution for the parameters of model m is highly diffuse, so that c_m is large, then (8) can be rewritten as

$$\log f(m|\boldsymbol{y}) \approx C + \log f(\boldsymbol{y}|m, \widehat{\boldsymbol{\beta}}_m) + \log f(m) - d_m \log c_m - \frac{d_m}{2} \log n \tag{10}$$

where $\widehat{\boldsymbol{\beta}}_m$ is the maximum likelihood estimate of $\boldsymbol{\beta}_m$. Equation (10) corresponds asymptotically to an information criterion with complexity penalty equal to $\log n + \log c_m^2 - 2d_m^{-1} \log f(m)$ compared with BIC, for example, where the complexity penalty is equal to $\log n$. The relative discrepancy between these two penalties is asymptotically zero. Poskitt and Tremayne (1983) discussed the interplay between prior model probabilities $f(m)$ and BIC and other information criteria in a time series context when Jeffreys priors are used for model parameters.

It is clear from (10) that a large value of c_m arising from a diffuse prior penalizes more complex models. On the other hand, a more moderate value of c_m (such as unit information) may have the effect of shrinking the posterior distributions of the model parameters towards the prior mean to a greater extent than desired. This has a particular impact when model averaging is used to provide predictive inferences (see, for example, Hoeting et al , 1999), where both the posterior model probabilities and the posterior distributions of the model parameters are important. A conflict can arise where to achieve the amount of dispersion desired in the prior distribution for model parameters, more complex models are unfairly penalized. To avoid this, we suggest choosing the dispersion of the prior distributions of model parameters to provide the amount of shrinkage to the prior mean which is considered appropriate a priori, and to choose prior model probabilities to adjust for the resulting effect this will have on the posterior model probabilities. We propose

$$f(m) \propto p(m) c_m^{d_m} \tag{11}$$

where $p(m)$ are baseline model probabilities which do not depend on the prior distributions of the model parameters, and might be expected not to depend on the dimensions of the models, although we do not prohibit this. With this choice of $f(m)$, (8) becomes

$$\begin{aligned} \log f(m|\boldsymbol{y}) &= C + \log f(\boldsymbol{y}|m, \widehat{\boldsymbol{\beta}}_m) - \frac{1}{2c_m^2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m) \\ &\quad + \log p(m) - \frac{d_m}{2}\log n + O_p(n^{-1/2}) \end{aligned} \tag{12}$$

Where the specification of the base variance Σ_m is not in terms of unit information, the extra term $-\log(|\Sigma_m||i(\boldsymbol{\beta}_m)|)/2$ is required in (12). When c_m^2 is large and when all $p(m)$ are equal, model comparison is asymptotically based on BIC. More generally, we propose choosing prior model probabilities based on (11) for any prior variance V_m. Substituting (9) into (11), we obtain

$$f(m) \propto p(m)(|V_m||i(\boldsymbol{\beta}_m)|)^{1/2}. \tag{13}$$

The choice of $p(m)$ can be based on the form of the equivalent model complexity penalty which is deemed to be appropriate *a priori*. Setting all $p(m)$ equal, which we propose as the default option, leads to model determination based on a modified BIC criterion involving penalized maximum likelihood. Hence, the impact of the prior distribution on the posterior model probability through $(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m)^T \Sigma_m^{-1} (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m)/2c_m^2$ in (12) is straightforward to assess, and any undesirable side effects of large prior variances are eliminated.

In order to specify prior model probabilities using (11), with $p(m)$ chosen to correspond to a particular complexity penalty, it is necessary to be able to evaluate c_m^{-2}, the number of units of information implied by the specified prior variance V_m for $\boldsymbol{\beta}_m$. Equivalently, as $f(m) \propto p(m)|V_m|^{\frac{1}{2}}|i(\boldsymbol{\beta}_m)|^{\frac{1}{2}}$, knowledge of $|i(\boldsymbol{\beta}_m)|$ is required. Except in certain circumstances, such as normal linear models, this quantity depends on the unknown model parameters $\boldsymbol{\beta}_m$. One possibility is to use a sample-based estimate $|i(\widehat{\boldsymbol{\beta}}_m)|$ to determine the 'prior' model probability, in which case the approach is not fully Bayesian. Alternatively, as suggested above, substituting $\boldsymbol{\mu}_m$, the prior mean of $\boldsymbol{\beta}_m$, into $i(\boldsymbol{\beta}_m)$ gives a prior for model comparison which has a unit information interpretation but for which model comparison is not asymptotically based on (12), the extra term $\log(|i(\boldsymbol{\mu}_m)|/|i(\boldsymbol{\beta}_m)|)/2$ being required.

3 Normal linear models

Here we consider normal linear models where for $m \in M$, $\boldsymbol{y} \sim N(\boldsymbol{X}_m\boldsymbol{\beta}_m, \sigma^2 I)$ with the conjugate prior specification

$$\boldsymbol{\beta}_m|\sigma^2, m \sim N(\boldsymbol{\mu}_m, \sigma^2 V_m) \quad \text{and} \quad \sigma^{-2} \sim \text{Gamma}(\alpha, \lambda)\ . \tag{14}$$

For such models the posterior model probabilities can be calculated exactly. Dropping the model subscript m for clarity,

$$f(m|\boldsymbol{y}) \propto f(m)\frac{|V^*|^{1/2}}{|V|^{1/2}}\left(2\lambda + \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\mu}^T V^{-1}\boldsymbol{\mu} - \widetilde{\boldsymbol{\beta}}^T(V^*)^{-1}\widetilde{\boldsymbol{\beta}}\right)^{-\alpha-n/2}$$

where $V^* = (V^{-1} + \boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\widetilde{\boldsymbol{\beta}} = V^*(V^{-1}\boldsymbol{\mu} + \boldsymbol{X}^T\boldsymbol{y})$ is the posterior mean. Hence, setting $V = c^2\Sigma$, as before,

$$\begin{aligned}
\log f(m|\boldsymbol{y}) &= C + \log f(m) - \frac{1}{2}\log|c^{-2}\Sigma^{-1} + \boldsymbol{X}^T\boldsymbol{X}| - \frac{1}{2}\log|\Sigma| - d\log c \\
&\quad - (\alpha + n/2)\log\left(2\lambda + \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\mu}^T V^{-1}\boldsymbol{\mu} - \widetilde{\boldsymbol{\beta}}^T(V^*)^{-1}\widetilde{\boldsymbol{\beta}}\right) \\
&= C - (\alpha + n/2)\log\left(2\lambda + (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\mu})^T V^{-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\mu})\right) \\
&\quad + \log f(m) - \frac{1}{2}\log|i| - \frac{d}{2}\log n - \frac{1}{2}\log|\Sigma| - d\log c + O(n^{-1})
\end{aligned} \tag{15}$$

where, with a slight abuse of notation, $i = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ is the unit information matrix multiplied by σ^2. Notice the correspondence between (7) and (15). As before, if $|\Sigma| = |i|^{-1}$, then c^{-2} can be interpreted as the number of units of information in the prior (as the prior variance is $c^2\sigma^2\Sigma$) and

$$\begin{aligned}
\log f(m|\boldsymbol{y}) &= C - (\alpha + n/2)\log\left(2\lambda + (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\mu})^T V^{-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\mu})\right) \\
&\quad + \log f(m) - \frac{d}{2}\log n - d\log c + O(n^{-1}).
\end{aligned} \tag{16}$$

In both (15) and (16) the posterior mean $\widetilde{\boldsymbol{\beta}}$ can be replaced by the least squares estimator $\widehat{\boldsymbol{\beta}}$. Again, if $c = 1$ (unit information) and the prior distribution across models is uniform, model comparison is performed using a modified version of BIC, as presented for example by Raftery (1995), where $n/2$ times the logarithm of the residual sum of squares for the model has been replaced by the first term on the right hand side of (16). The residual sum of squares is evaluated at the posterior mode, and is penalised by a term representing deviation from the prior mean, as in (7). This expression also depends on the prior for σ^2 through the prior parameters α and λ, although these terms vanish when the improper prior $f(\sigma^2) \propto \sigma^{-2}$, for which $\alpha = \lambda = 0$, is used. With these values, and setting $\Sigma^{-1} = i = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, we obtain the prior used by Fernandez et al (2001), who also note the unit information interpretation when $c = 1$.

As before, if the prior variance V suggests a different value of c, then the resulting impact on the posterior model probabilities can be moderated by an appropriate choice of $f(m)$ and again we propose the use of (11) and (13), noting that for normal models i is known. In the context of normal linear models, Pericchi (1984) suggests a similar adjustment of prior model probabilities by an amount related to the expected gain in information. Alternatively, replacing $|i|$ by $|i+n^{-1}V^{-1}|$ in (13), resulting in

$$f(m) \propto p(m)|V|^{\frac{1}{2}}|i+n^{-1}V^{-1}|^{\frac{1}{2}}, \tag{17}$$

makes (15) exact, eliminating the $O(n^{-1})$ term. Again, for highly diffuse prior distributions on the model parameters (large values of c^2), together with $\alpha=\lambda=0$ and prior model probabilities based on (11) and (13), equation (16) implies that model comparison is performed on the basis of BIC.

4 Relationship with other information criteria

In Sections 2 and 3, we have investigated how prior model probabilities might be specified by considering their joint impact, together with the prior distributions for the model parameters, on the posterior model probabilities. It was shown that making these probabilities depend on the prior variance of the associated model parameters using (11) or (13) with uniform $p(m)$ leads to posterior model probabilities which are asymptotically equivalent (to order $n^{-\frac{1}{2}}$) to those implied by BIC. For models other than normal linear regression models, a prior value of $\boldsymbol{\beta}$ must be substituted into (13) and so the approximation only attains this accuracy for $\boldsymbol{\beta}$ within an $O(n^{-\frac{1}{2}})$ neighbourhood of this value. Nevertheless, we might expect BIC to more accurately reflect the full Bayesian analysis for such a prior than more generally, where the error of BIC as an approximation to the log-Bayes factor is $O(1)$.

Alternative (non-uniform) specifications for $p(m)$ might be based on other information criteria of the form

$$\log f(\boldsymbol{y}|m,\widehat{\boldsymbol{\beta}}_m)-\frac{1}{2}\psi(n)d_m$$

where $\psi(n)$ is a 'penalty' function; for BIC, $\psi(n)=\log n$ and for AIC $\psi(n)=2$. From (12), for large c_m^2 or for a modified criterion, we have $\psi(n)=\log n+2d_m^{-1}\log p(m)$. As $p(m)$ contributes to the prior model probability through (11) it cannot be a function of n since our prior belief on models should not change as the sample size changes. Therefore, strictly, the only penalty functions which can be equivalent to setting prior model probabilities as in (11) are of the form $\psi(n)=\log n+\psi_0$ for some positive constant $\psi_0>0$. Any alternative dependence on n would correspond to a prior which depended on n, through $f(m)$ or $f(\boldsymbol{\beta}_m|m)$. Hence AIC, for example, is prohibited (as would be expected, as AIC is not consistent, whereas any approach arising from a proper prior must be). Nevertheless, if a penalty function of a particular form is desired for a sample of a specified size n_0, then setting $\log p(m)=\frac{d_m}{2}\Big\{\log n_0-\psi(n_0)\Big\}$ will ensure that posterior model probabilities are calculated on the basis of the information criterion with penalty $\psi(n_0)$, at the relevant sample size n_0.

5 Alternative arguments for $f(m) \propto c_m^{d_m}$

The strategy described in this paper can be viewed as a full Bayesian approach where the prior distribution for model parameters is specified by focusing on the uncertainty concerning those parameters alone, and the prior model probabilities can be specified by considering the way in which an associated 'information criterion' balances parsimony and goodness-of-fit. In the past, informative specifications for these probabilities have largely been elicited via the notion of imaginary data; see for example Chen et al (1999, 2003). Within the approach suggested here, prior model probabilities are specified by considering the way in which data yet to be observed might modify ones beliefs about models, given the prior distributions for the model parameters. Full posterior inference under model uncertainty, including model averaging, is then available for the chosen prior.

Specifying the prior distribution on the basis of how it is likely to impact the posterior distribution is entirely valid, but may perhaps seem unnatural. In particular, the consequence that the prior model probabilities might depend on the prior distributions for the model parameters may seem somewhat alien. This is particularly true of the implication of (13), that models where we have more information (smaller dispersion) in the prior distribution should be given lower prior probabilities than models for which we are less certain about the parameter values. One justification for this is to examine the prior model probabilities for particular subsets of the parameter spaces within models. This can be considered as an extension of the approach of Robert(1993) for two normal models. We consider the prior probability of the event

$$E = \{\text{model } m \text{ is 'true' }\} \cap \{(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m)^T i(\boldsymbol{\beta}_m^0)(\boldsymbol{\beta}_m - \boldsymbol{\mu}_m) < \epsilon^2\}$$

for some reference parameter value $\boldsymbol{\beta}_m^0$, possibly the prior mean $\boldsymbol{\mu}$. The dependence of this subset of the parameter space on the unit information at $\boldsymbol{\beta}_m^0$ enforces some degree of comparability across models. This is particularly true if the various values of $\boldsymbol{\beta}_m^0$ are compatible (for example they imply the same linear predictor in a generalised linear model, as they would generally do if set equal to $\mathbf{0}$). For the purposes of the current discussion, we also require $V_m = c_m^2 i(\boldsymbol{\beta}_m^0)^{-1}$. This is a plausible default choice, but nevertheless represents considerable restriction on the structure of the prior variance, which was previously unconstrained. Then

$$\begin{aligned} P(E) &= f(m)P\left(\chi^2_{d_m} < \frac{\epsilon^2}{c_m^2}\right) \\ &\approx \frac{f(m)\epsilon^{d_m}}{2^{d_m/2-1}\Gamma(d_m/2)c_m^{d_m}} \end{aligned}$$

for small ϵ. Therefore, for this prior, if the joint prior probability of model m in conjunction with $\boldsymbol{\beta}_m$ being in some specified neighbourhood (defined according to a unit information inner product) of its prior mean is to be uniform across models then we require $f(m) \propto p(m)c_m^{d_m}$ as in (11), with $p(m) = 2^{d_m/2-1}\Gamma(d_m/2)/\epsilon^{d_m}$.

An alternative justification of (11) when the model parameters are given diffuse normal prior distributions arises as follows. One way of taking a 'baseline' prior distribution and making it more diffuse, to represent greater prior uncertainty, is to raise the prior density to the power $1/c^2$ for some $c^2 > 1$, and then renormalise. For example, for a single normal distribution this has the effect of multiplying the variance by c^2, which increases the prior dispersion in an obvious way. Highly diffuse priors, suitable in the absence of strong prior information, may be thought as arising from a baseline prior transformed in this way for some large value of c^2. Where model uncertainty exists, the joint prior distribution is a mixture whose components correspond to the models, with mixture weights $f(m)$. As suggested above, a diffuse prior distribution might be obtained by raising a baseline prior density (with respect to the natural measure over models and associated parameter spaces) to the power $1/c^2$ and renormalising. Where the baseline prior distribution for $\boldsymbol{\beta}_m$ is normal with mean $\boldsymbol{\mu}_m$ and variance Σ_m, the effect of raising the mixture prior density to the power $1/c^2$ is to increase the variance of each $\boldsymbol{\beta}_m$ by a factor of c^2, as before. For large values of c^2 the effect of the subsequent renormalisation is that the model probabilities are proportional to $|\Sigma_m|^{1/2}(2\pi)^{d_m/2}c^{d_m}$, independent of the model probabilities in the original baseline mixture prior. Again this illustrates a relationship between prior model probabilities and prior dispersion parameters satisfying (11). For the two normal models considered by Robert (1993) the resulting prior model probabilities are identical. Where the baseline variance is based on unit information, so $|\Sigma_m| = |i(\boldsymbol{\beta}_m)|$, then the prior model probabilities can be written as (13) with $p(m) = (2\pi)^{d_m/2}|i(\boldsymbol{\beta}_m)|^{-1/2}$.

Finally, this approach can be justified by considering the behaviour of the posterior mean under model averaging. We restrict consideration here to two nested models, m_0 and m_1, differing by a single parameter θ and suppose that $f(y|m_0) = f(y|m_1, \theta_0)$. We assume that the (marginal) prior for θ under m_1 is $N(\theta_0, \tau^{-1})$ and, without loss of generality, we take $\theta_0 = 0$. Under model m_1 the Bayes estimator for θ is the posterior mean $E_1(\theta|y)$, which has asymptotic expansion

$$E_1(\theta|y) = \widehat{\theta}\left(1 - \frac{i(\widehat{\theta})\tau}{n}\right) + \frac{a_3}{2i(\widehat{\theta})^2 n} + o(n^{-1}) \tag{18}$$

where na_3 is the third derivative of the log-likelihood, evaluated at $\widehat{\theta}$ (see for example, Johnson, 1970, Ghosh, 1994). This illustrates the usual effect of prior precision τ as a shrinkage parameter, with the posterior mean being shrunk away from the m.l.e., with the amount of shrinkage diminishing as $\tau \to 0$. Hence, for fixed y, the posterior mean for θ is (asymptotically) monotonic in τ. Allowing for model uncertainty, we have $E(\theta|y) = f(m_1|y)E_1(\theta|y)$ where

$$f(m_1|y) = \frac{1}{1 + k(2\pi)^{1/2}\tau^{-1/2}f_1(0|y)} \tag{19}$$

where $f_1(\theta|y)$ is the posterior (marginal) density for θ under m_1, and k are the prior odds in favour of m_0 over m_1. Combining (18) and (19), we see that the relationship between the coefficient for $\widehat{\theta}$ in the model averaged posterior depends and the prior precision for θ is no longer generally monotonic, so τ no longer has a simple interpretation as a shrinkage parameter. A simple illustration of this is provided by Figure 1, where this coefficient is plotted for various values of τ, for the simple example of a normal distribution with known error variance, and prior odds $k = 1$. It can be seen that, regardless of the value of τ there will be a certain amount of shrinkage to the prior mean. Adopting the approach advocated in this paper has the effect of setting $k \propto \tau^{1/2}$ which mitigates this effect, and returns control over the shrinkage to the analyst.

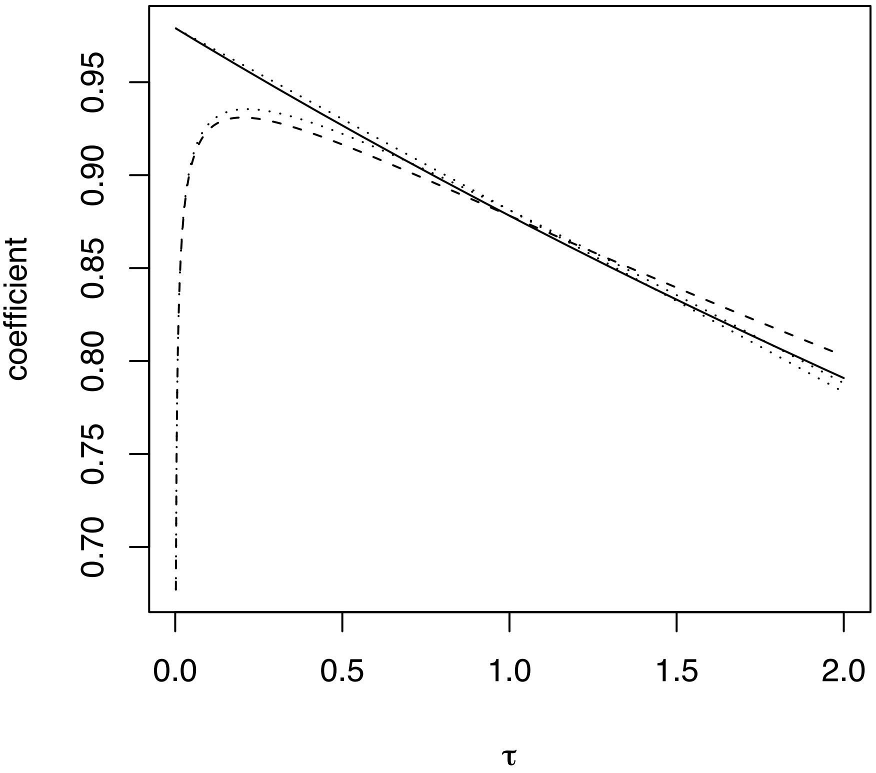


Figure 1: Model average coefficient on $\widehat{\theta}$ [evaluated as $\widehat{\theta}/\theta$], for normal likelihood with known error variance, σ^2. The plot here is for $n = 10$, $\widehat{\theta} = 1$, $\sigma^2 = 1$. The solid line is for a uniform prior over models, and the dashed line uses prior model probability $f(m_1) \propto \tau^{1/2}$. The dotted lines are approximations based on replacing $(2\pi)^{1/2} f_1(0|y)$ in (19) with its normal approximation $\exp\left(-\frac{i(\widehat{\theta})n}{2}\widehat{\theta}^2\right)$, ignoring the dependence, to $O(n^{-1})$, of $f_1(0|y)$ on τ.

The purpose of the above discussion is not necessarily to advocate a particular prior, but simply to illustrate that one can arrive at (11) by direct consideration of prior probabilities, or prior densities, or by the behaviour of posterior means, as well as by the asymptotic behavior of posterior model probabilities, or associated numerical approximations, as earlier.

6 Illustrated Examples

6.1 Scope

Here we present three examples. In Section 6.1 we illustrate the effect of Lindley's paradox in a standard linear regression context emphasizing its dramatic effect on inference concerning model uncertainty. At the same time, we demonstrate that if instead of using the standard discrete uniform prior distribution (DU) for $f(m)$ we adopt our proposed discrete adjusted prior distribution (DA) given by (11) with $p(m) = 1$, this effect is diminished.

Section 6.2 illustrates that unit information prior specifications (or other specifications suggesting smaller prior parameter dispersion) can indeed significantly shrink posterior distributions towards zero. This effect suggests that although prior variances based on unit information might have desirable behaviour with respect to model determination, they may unintentionally distort the parameter posterior distributions. We demonstrate that this can affect the predictive ability of routinely used model averaging approaches in which information is borrowed across a set of models.

Finally, Section 6.3 investigates the behaviour of posterior model probabilities when substantive prior information about the parameters ia available. We demonstrate through a real data example that the DU prior may have a significant impact on posterior model probabilities and we illustrate the advantages of choosing the DA prior model probabilities that are appropriately adjusted for parameter prior dispersions.

6.2 Example 1: Simulated Regression Example

We consider a simulated dataset based on $n = 50$ observations of 15 standardized normal covariates $X_j,\ j = 1, \ldots, 15$, and a response variable Y generated as

$$Y \sim N(\ X_4 + X_5,\ 2.5^2\ )\ .$$

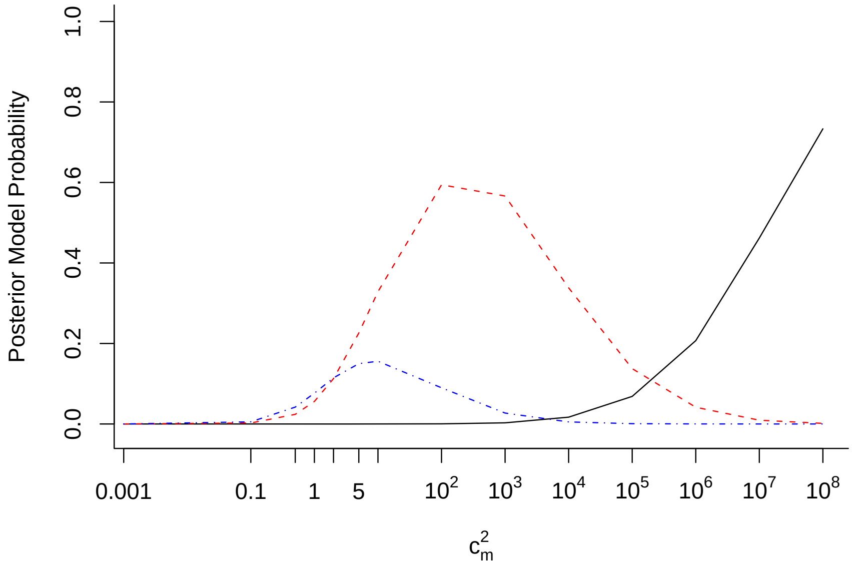
Assuming a conjugate normal inverse gamma prior distribution given by (14) with zero mean, $V_m = c_m^2 \Sigma_m$ and $a = \lambda = 10^{-2}$, we calculated posterior model probabilities for all models under consideration. Similar behaviour is exhibited either when Σ_m is specified as $\Sigma_m = n \left(\boldsymbol{X}_m^T \boldsymbol{X}_m \right)^{-1}$ (described below) or as $\Sigma_m = \boldsymbol{I}_{d_m}$.

Figure 2(a) illustrates Lindley's paradox for this dataset with DU prior. Simpler models are preferred as c_m^2 increases. In contrast, the DA prior in Figure 2(b) identifies $1+X_4+X_5+X_{12}$ as the highest probability model for any value of $c_m^2 > 1$. Note that, when $\Sigma_m = n \left(\boldsymbol{X}_m^T \boldsymbol{X}_m \right)^{-1}$, $c_m^2 = 1$ represents the dispersion induced by the unit information prior. Similarly, Figure 3 summarises the posterior inclusion probability of each variable X_j. Again, in for the DU prior these probabilities are sensitive to changed in c_m^2 across its range, whereas the DA prior produces stable results for $c_m^2 > 1$.

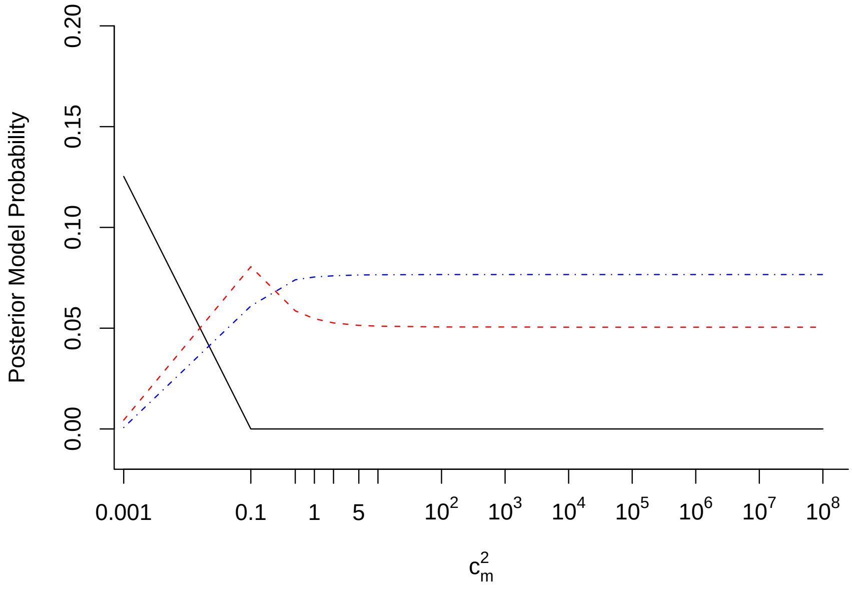
6.3 Example 2: A real data linear regression example

Montgomery et al (2001) investigate the effect of the logarithm of wind velocity (x), measured in miles per hour, on the production of electricity from a water mill (y), measured in volts, via a linear regression model of the form

$$y_i \sim N\Big(\beta_0 + \beta_1 x_i, \sigma^2\Big),\ i = 1, \ldots, n$$

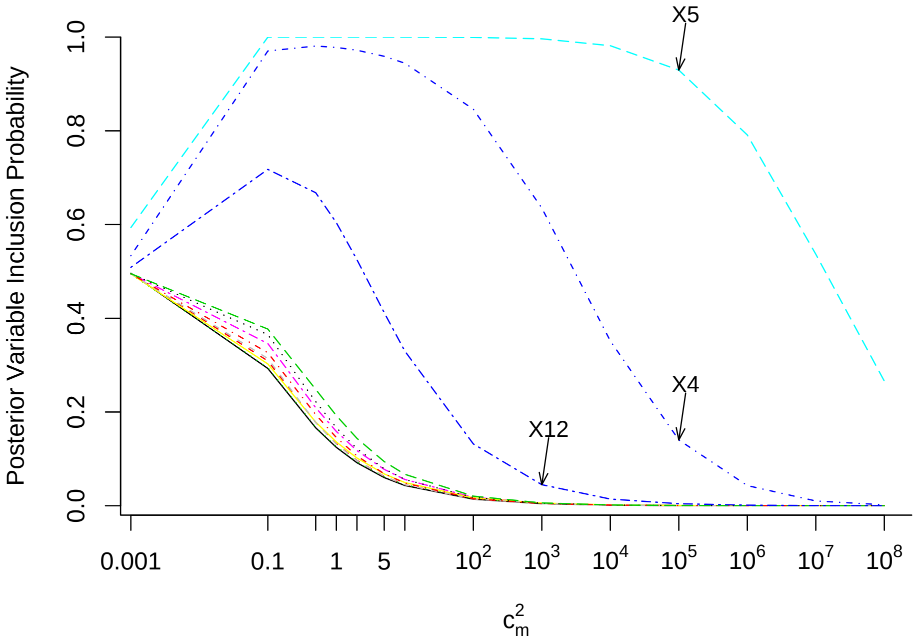


(a) Discrete Uniform (DU) prior

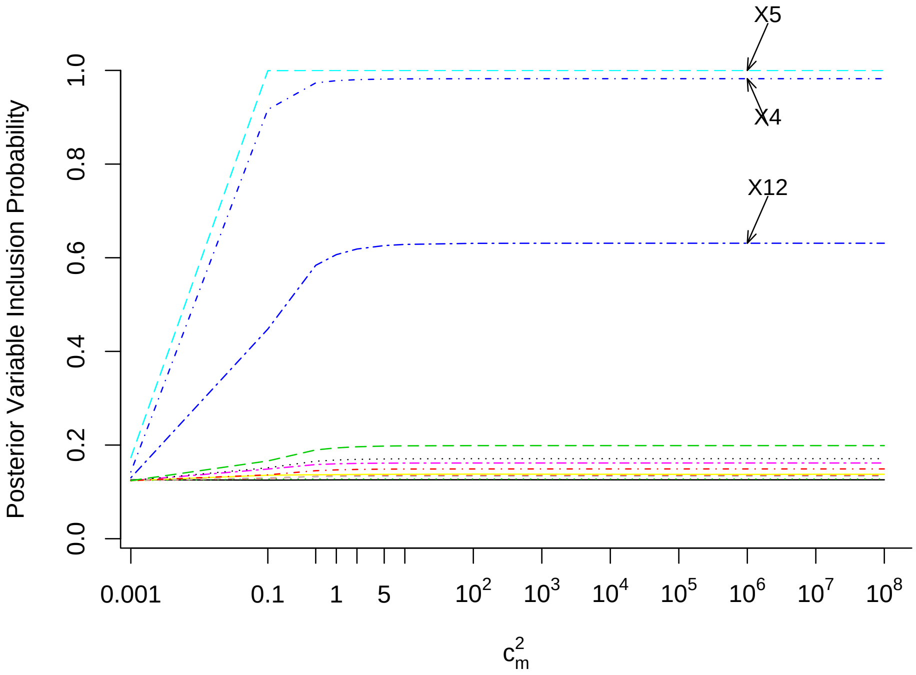


(b) Adjusted discrete (DA) prior

Figure 2: Posterior model probabilities under different prior dispersions. Solid line: constant model; short dashed line: $1 + X_4 + X_5$ model; long dashed line: $1 + X_4 + X_5 + X_{12}$ model.



(a) Discrete Uniform (DU) prior



(b) Adjusted discrete (DA) prior

Figure 3: Posterior variable inclusion probabilities under different prior dispersions.

based on $n = 25$ data points. We calculate the posterior odds of the above model, denoted by m_1, against the constant model denoted by m_0, adopting the usual conjugate prior specification given by (14) with zero mean, variance $V_m = c_m^2 n \left(\boldsymbol{X}_m^T \boldsymbol{X}_m\right)^{-1}$ and $\alpha = \lambda = 10^{-2}$. Since there is a high sample correlation coefficient of 0.978 between y and x, we expect that m_1 will be a posteriori strongly preferred to m_0. Indeed, the posterior probability of m_1 is very close to one for values of c_m^2 as large as 10^{28}. This behaviour provides a source of security with respect to the choice of c_m^2 and Lindley's paradox, but we should also investigate the effect of c_m^2 on the posterior densities of β_0 and β_1; see Figure 4. We have used values of c_m^2 that represent highly diffuse priors with $c_m^2 = 10$ and $c_m^2 = 100$, the unit information prior that approximates BIC with $c_m^2 = 1$, a prior that approximates AIC for this sample size $c_m^2 = (e^2 - 1)/n = 0.256$ and a prior suggested by the risk inflation criterion (RIC) of Foster and George (1994) with $c_m^2 = 0.04$. It is striking that the resulting posterior densities differ highly in both location and scale. The danger of misinformation when unit information priors are used is discussed in detail by Paciorek (2006). The approach described in this paper allows, where considered appropriate, the prior distribution for the model parameters to be made highly diffuse, so that it does not impact strongly on the posterior model parameters, while at the same time, through a DA prior across models, ensuring that posterior model probabilities are unduly skewed.

We now investigate the effect of prior specification when prediction is of primary interest. Assume that predictions will be based on the MCMC output estimate of the model-averaging predictive density of observation y_i given the rest of the data $\boldsymbol{y}_{\setminus i}$,

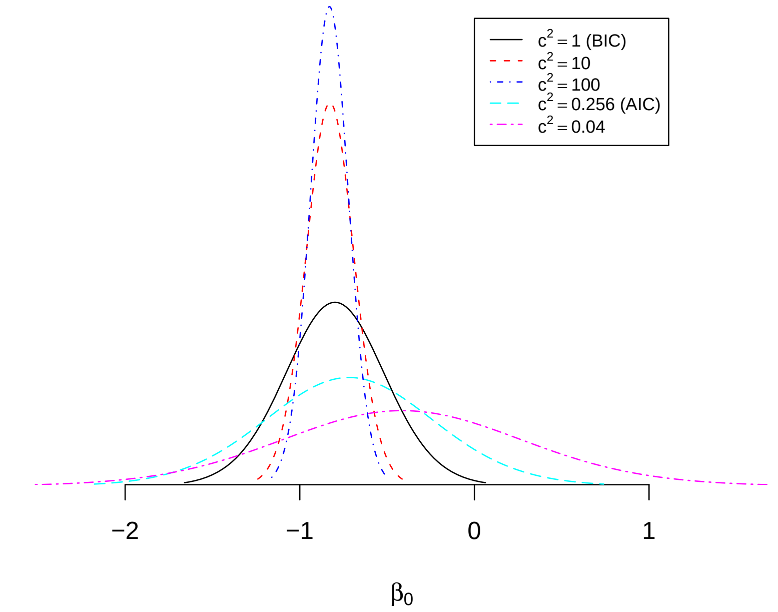
$$f^p(i) = \sum_{m \in M} f(m) f(y_i | \boldsymbol{y}_{\setminus i}, m).$$

To evaluate the predictive performance, as a function of the prior, we apply the negative cross-validatory log-likelihood score (NCV; see Geisser and Eddy, 1979) given by

$$NCV = -\sum_{i=1}^{n} \log f^p(i).$$

Lower values of NCV indicate greater predictive accuracy. Following Gelfand (1996) we estimate $f^p(i)$ by the inverse of the posterior (over $m, \boldsymbol{\beta}_m$) mean of the inverse predictive density of observation i.

We generated three additional covariates that have correlation coefficients 0.99, 0.97 and 0.89 with x and performed the same model determination exercise. Posterior model probabilities for all models were calculated for all models under consideration. We used Zellner's g-prior $V_m = c_m^2 n \left(\boldsymbol{X}_m^T \boldsymbol{X}_m\right)^{-1}$ and an independence prior $V_m = c_m^2 \boldsymbol{I}_{d_m}$. For the DU prior combined with the unit information prior obtained by $c_m^2 = 1$, NCV is far away from the minimum value achieved for higher values of c_m^2; see Figure 5(a). For $c_m^2 > 10^5$ NCV increases due to the effect of Lindley's paradox focusing posterior probability on models that are unrealistically simple. On the other hand, our proposed DA prior specification achieves the maximum predictive ability for any large values of c_m^2; see Figure 5(b).



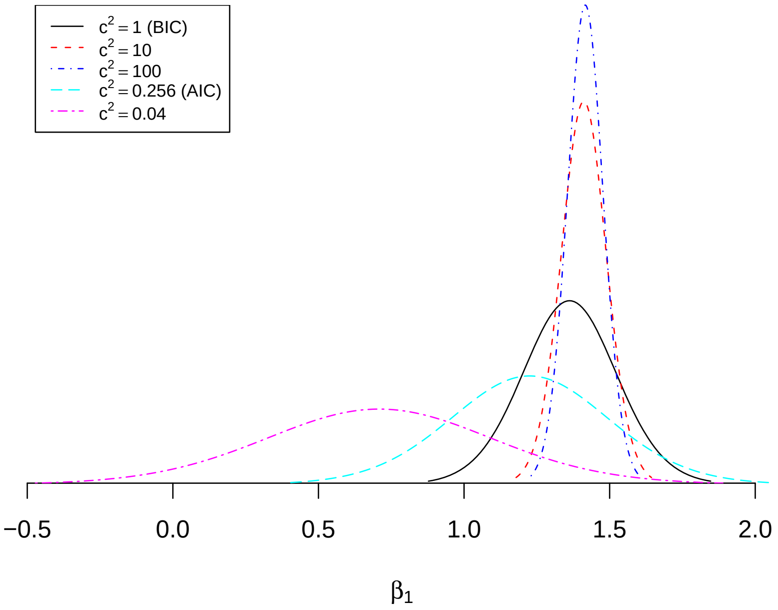
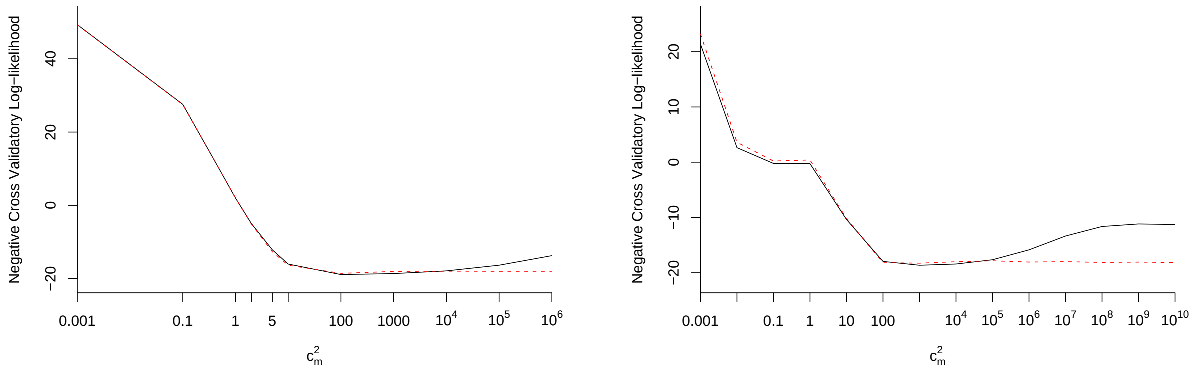


Figure 4: Posterior densities of parameters β_0 and β_1 under different prior dispersions; $c_m^2 = c^2$ for all models m.

This simulated data exercise does indicate that predictive ability can be optimised if highly dispersed prior parameter densities are chosen together with the DA prior over model space.



(a) Zellner's g-prior $\left(V_m = c_m^2 n \left(\boldsymbol{X}_m^T \boldsymbol{X}_m\right)^{-1}\right)$ (b) Independence prior $(V_m = c_m^2 \boldsymbol{I}_{d_m})$

Figure 5: Negative cross-validatory log-likelihood for two prior dispersion structures with DU prior (solid line) and DA prior (dashed line).

6.4 Example 3: $3 \times 2 \times 4$ Contingency Table Example with Available Prior Information

We consider data presented by Knuiman and Speed (1988) to illustrate how our proposed methodology performs in an example where prior information for the model parameters is available. The data consist of 491 individuals classified in n cells by categorical variables obesity (O: low,average,high), hypertension (H: yes,no) and alcohol consumption (A: 1,1–2,3–5,6+ drinks per day). We adopt the notation of the full hierarchical log-linear model used by Dellaportas and Forster (1999)

$$y_i \sim Poisson(\lambda_i) \text{ for } \ i = 1, 2, \ldots, n, \quad \log(\boldsymbol{\lambda}) = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$, $\boldsymbol{X}$ is the $n \times n$ design matrix of the full model $\boldsymbol{\beta} = (\boldsymbol{\beta}_j; j \in \mathcal{V})$ is a $n \times 1$ parameter vector, $\boldsymbol{\beta}_j$ are the model parameters that correspond to j term and $\mathcal{V}$ is the set of all terms under consideration. All parameters here are defined using the sum-to-zero constraints. Dellaportas and Forster (1999) proposed as a default prior for parameters of log-linear models

$$\boldsymbol{\beta}_j \sim N\left(\boldsymbol{\mu}_j, \quad k_j^2 \left(\boldsymbol{X}_j^T \boldsymbol{X}_j\right)^{-1}\right) \tag{20}$$

with $\boldsymbol{\mu}_j$ being a vector of zeros and $k_j^2 = 2n$ for all $j \in \mathcal{V} = \{\emptyset, O, H, A, OH, OA, HA, OHA\}$; we denote this prior by DF.

In their analysis, Knuiman and Speed (1988) took into account some prior information available about the parameters $\boldsymbol{\beta}_j$. In particular, prior to this study information was available indicating that $\boldsymbol{\beta}_{OHA}$ and $\boldsymbol{\beta}_{OA}$ are negligible and only $\mathcal{V} = \{\emptyset, O, H, A, OH, HA\}$ should be considered. Moreover, the term $\boldsymbol{\beta}_{HA}$ is non-zero with a-priori estimated effects $\overline{\boldsymbol{\beta}}_{HA}^T = (0.204, -0.088, -0.271)$; (note that the signs of the prior mean are opposite when compared with reported values of Knuiman and Speed since we have used a different ordering of the variable levels).

Knuiman and Speed adopted the prior (20) with $\boldsymbol{\mu}_{HA} = \overline{\boldsymbol{\beta}}_{HA}$ and $\boldsymbol{\mu}_j = \mathbf{0}$ for $j \in \mathcal{V} \setminus \{HA\}$ and prior variance coefficients $k_{HA}^2 = 0.05$ and $k_j^2 = \infty$ for $j \in \{\emptyset, O, H, A, OH\}$. In our data analysis we used $k_j^2 = 10^4$ instead of $k_j^2 = \infty$. We denote this prior as KS. We also used a combination of the DF and KS priors, denoted by KS/DF, modifying slightly the KS prior so that $k_j^2 = 2n$ for terms $j \in \{\emptyset, O, H, A, OH\}$. Finally, an additional diffuse independence prior, denoted by IND, with zero prior mean and variance 10^3 for all model parameters was also used.

In log-linear models $i(\boldsymbol{\beta}_m)$ depends on $\boldsymbol{\beta}_m$ so to specify the DA prior we utilize the prior mean $\boldsymbol{\mu}_m$ of $\boldsymbol{\beta}_m$ resulting in

$$f(m) \propto p(m)|V_m|^{1/2}|\boldsymbol{X}_m^T Diag(\boldsymbol{\lambda}_0)\boldsymbol{X}_m|^{1/2} n^{-d_m/2}, \quad \boldsymbol{\lambda}_0 = \exp\left(\boldsymbol{X}_m \boldsymbol{\mu}_m\right) \ ,$$

while the prior parameters $p(m)$ were set equal to $\log p(m) = -\frac{d_m}{2}\log(2)$ in line with the DF prior.

Parameter	Model space	Prior model probabilities				Posterior model probabilities			
Prior	Prior	O+H+A	OH+A	O+HA	OH+HA	O+H+A	OH+A	O+HA	OH+HA
1. DF	DU	0.25	0.25	0.25	0.25	0.657	0.336	0.004	0.002
2. KS	DU	0.25	0.25	0.25	0.25	0.075	0.000	0.923	0.002
3. KS/DF	DU	0.25	0.25	0.25	0.25	0.059	0.023	0.638	0.280
4. DF	DA	0.247	0.247	0.251	0.255744	0.677	0.317	0.004	0.002
5. KS	DA	0.046	0.954	2.0×10^{-6}	3.3×10^{-5}	0.665	0.335	0.000	0.000
6. KS/DF	DA	0.500	0.500	1.7×10^{-5}	1.7×10^{-5}	0.690	0.310	0.000	0.000
7. IND	DA	0.003	0.996	3.0×10^{-6}	0.001	0.690	0.303	0.004	0.003

Table 1: Prior and posterior model probabilities under different parameter and model prior densities.

Posterior model probabilities (estimated using RJMCMC) for all prior specifications are presented in Table 1. The top right of the Table illustrates the striking effect of informative parameter priors on posterior model probabilities. The difficulty to make joint inferences on the product parameter and model space is evident by inspecting the sensitivity of model probabilities to different prios. However, the DA specification adjusts the prior model probabilities so that posterior model probabilities are robust under all prior specifications.

7 Conclusion

There are clearly alternative specifications for the prior model probabilities $p(m)$ which satisfy (11), and we do not seek to justify one over the other. Indeed, choosing model probabilities to satisfy (11) may not be appropriate in some situations. Hence, we do not propose (11) as a necessary condition for $f(m)$ although we do believe that there are compelling reasons for considering such a specification, perhaps as a default or reference position in the type of situations we have considered in this paper. What we do argue is that there is nothing sacred about a uniform prior distribution over models, and hence by implication, about the Bayes factor. It is completely reasonable to consider specifying $f(m)$ in a way which takes account of the prior distributions for the model parameters for individual models. Then, certainly within the contexts discussed in this paper, as demonstrated by the examples we have presented, the issues surrounding the role of the prior distribution for model parameters, in examples with model uncertainty, become much less significant.

References

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

[2] Chen, M. H. , Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models. *Journal of the Royal Statistical Society B*, **61**, 223–243

[3] Chen, M. H. , Ibrahim, J. G. Shao, Q. -M. and Weiss, R. E. (2003). Prior Elicitation for Model Selection and Estimation in Generalized Linear Mixed Models. *Journal of Statistical Planning and Inference*, **111**, 57 – 76.

[4] Chipman, H. (1996). Bayesian Variable Selection with Related Predictors.*Canadian Journal of Statistics*, **24**, 17–36.

[5] Chipman H., George E.I. and McCulloch R.E. (2001) *The practical implementation of Bayesian Model Selection.* IMS Lecture Notes - Monograph Series **38**.

[6] Dawid A. P. (1999) The Trouble with Bayes Factors. *Research report 202*, Department of Statistical Science, University College London, UK.

[7] Dellaportas, P. and Forster, J. J. (1999). Markov Chain Monte Carlo Model Determination for Hierarchical and graphical log-linear Models. *Biometrika*, **86**, 615–633.

[8] Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* New York, Wiley.

[9] Fernandez, C. , Ley, E. and Steel, M. F. J. (2000). Benchmark Priors For Bayesian Model Averaging. *Journal of Econometrics*, **100**, 381–427.

[10] Foster, D.P. and George, E.I. (1994) The risk inflation criterion for multiple regression. *Annals of Statistics*, **22**, 1947-1975.

[11] Geisser, S. and Eddy, W.F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, **74**, 153–160, (*Corrigenda in vol. 75, p. 765*).

[12] Gelfand, A.E. (1996). Model Determination Using Sampling-Based Methods. *Markov Chain Monte Carlo in Practice*, (eds. Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.), Chapman & Hall, Suffolk, UK, 145–161.

[13] Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, 711–732.

[14] Green, P. J. (2003). Trans-dimension al Markov Chain Monte Carlo (with discussion). In *Highly Structured Stochastic Systems.* (Green, P. J. , Hjort, N. L. and Richardson, S. , eds.) Oxford: Oxford University Press, 179-198.

[15] Hans, C., Dobra, A. and West, M. (2007) Shotgun Stochastic Search for "large p" regression. *Journal of the American Statistical Association*, **102**, 507–517.

[16] Hoeting, J. A. , Madigan, D. , Raftery, A. E. and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial (with discussion). *Statistical Science*, **14**, 382–417.

[17] Jeffreys, H. (1961). *Theory of Probability.* 3rd Edition, Cambridge. MA; New York: Oxford University Press.

[18] Kadane, J.B. and Lazar A. (2004). Methods and Criteria for Model Selection. *Journal of the American Statistical Association*, **99**, 279–290.

[19] Kass, R. E., Tierney, L. and Kadane, J.B. (1988). Asymptotics in Bayesian computation. In *Bayesian Statistics 3.* (Bernardo, J.M. et al, eds.) Oxford: Oxford University Press, 261-274.

[20] Kass, R. E. and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, **90**, 928–934.

[21] Knuiman, M.W. and Speed, T.P. (1988). Incorporating Prior Information Into the Analysis of Contingency Tables. *Biometrics*, **44**, 1061–1071.

[22] Kohn, R., Smith, M. and Chan, D. (2001) Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, **11**, 313-322.

[23] Laud, P.W. and Ibrahim, J.G. (1995).Predictive Model Selection.*Journal of the Royal Statistical Society B*, **57**, 247–262.

[24] Laud, P.W. and Ibrahim, J.G. (1996).Predictive Specification of Prior Model Probabilities in Variable Selection.*Biometrika*, **83**, 267–274.

[25] Liang, F., Paulo, R., Molina, G., Clyde, M. A.and Berger J. O. (2007) Mixtures of g-prior for Bayesian Variable Selection, *Journal of the American Statistical Association*, to appear.

[26] Lindley, D. V. (1957). A Statistical Paradox. *Biometrika*, **44**, 187–192.

[27] Madigan, D., Raftery, A.E., York, J., Bradshaw, J.M. and Almond, R.G. (1995).Strategies for Graphical Model Selection.*Selecting Models from Data: AI and Statistics IV* (P. Cheesman and R. W. Oldford, eds.). Berlin: Springer Verlag, 91–100.

[28] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*. Third edition, Wiley, New York.

[29] Ntzoufras, I. , Dellaportas P. and Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of statistical planning and inference*, **111**, 165-180.

[30] Paciorek C. J. (2006). Misinformation in the conjugate prior for the linear model with implications for free-knot spline modelling. *Bayesian Analysis*, **1**, 375–383.

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

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

[33] Raftery, A. E. (1995). Bayesian Model Selection for Social Research (with discussion). *Sociological Methodology 1995*. P. V Marsden, ed. 111–196. Oxford: Blackwell.

[34] Raftery, A. E. (1996). Approximate Bayes Factors and Accounting for model uncertainty in Generalized Linear Models. *Biometrika*, **83**, 251–266.

[35] Robert, C. P. (1993). A Note on Jeffreys-Lindley Paradox. *Statistica Sinica*, **3**, 601-608.

[36] Schervish, M. J. (1995). *Theory of Statistics*. 2nd edition. Springer Verlag.

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

[38] Volinsky C. T. and Raftery A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, **56**, 256-262.

[39] Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. Bayesian Statistics, **1**: Proceedings of the First International Meeting held in Valencia (Spain), (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), Valencia: University Press, pp. 585–603.