Statistical Science
2012, Vol. 27, No. 2, 232–246
DOI: 10.1214/11-STS369

Joint Specification of Model Space and Parameter Space Prior Distributions

Petros Dellaportas, Jonathan J. Forster and Ioannis Ntzoufras

Abstract. 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.

Key words and phrases: Bayesian inference, BIC, generalized 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 $\mathbf{y}$ are considered to have been generated by a model m, one of a set M of competing models. Each model specifies the distribution of $\mathbf{Y}$, $f(\mathbf{y}|m, \boldsymbol{\beta}_m)$ 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

$$(1) \qquad f(m|\mathbf{y}) = \frac{f(m) f(\mathbf{y}|m)}{\sum_{m \in M} f(m) f(\mathbf{y}|m)}, \qquad m \in M,$$

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

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

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

$$(2) \qquad \frac{f(m_1|\mathbf{y})}{f(m_2|\mathbf{y})} = \frac{f(m_1)}{f(m_2)} \frac{f(\mathbf{y}|m_1)}{f(\mathbf{y}|m_2)},$$

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

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

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* $\mathbf{y}$, the Bayes factor in favor of m_1 can be made arbitrarily large by choosing a sufficiently diffuse prior distribution for $\boldsymbol{\beta}_{m_2}$ (corresponding to a prior

Petros Dellaportas is Professor, Department of Statistics, Athens University of Economics and Business, 10434 Athens, Greece (e-mail: petros@aueb.gr). Jonathan J. Forster is Professor, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, UK (e-mail: J.J.Forster@soton.ac.uk). Ioannis Ntzoufras is Associate Professor, Department of Statistics, Athens University of Economics and Business, 10434 Athens, Greece (e-mail: ntzoufras@aueb.gr).

density $f(\boldsymbol{\beta}_{m_2}|m_2)$ which is sufficiently small over the range of values of $\boldsymbol{\beta}_{m_2}$ with nonnegligible 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 (2011) pointed out, the Bayes factor is only one of the two elements on the right side of (2) which contribute toward 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 nonuniform prior distributions have been suggested include the works of Madigan et al. (1995), Chipman (1996), Laud and Ibrahim (1995, 1996), Chipman, George and McCulloch (2001), Cui and George (2008), Ley and Steel (2009) and Wilson et al. (2010). We propose a different approach where 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. This leads to a nonuniform specification for the prior distribution over models, depending directly on the prior distributions for model parameters.

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 (2011) proposed 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 suggested an approach for constructing a general noninformative 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. Clyde (2000) proposed a similar approach where the priors for parameters of individual models are uniform and the relative weights of different models are chosen by constraining the resulting posterior model probabilities to be equivalent to those resulting from a specified information criterion, such as BIC.

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 under transformation) so 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 $|V|^{-1/2}g((\boldsymbol{\beta}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}))$ 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. Similarly, our approach can also be adapted to common prior distributions for parameters of graphical models.

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 $|\Sigma_m|$, although for the remainder of this section we do not require an explicit value; further discussion of this issue is presented in Section 2. Hence, suppose that

$$\begin{aligned} &f(\boldsymbol{\beta}_m|m)\\ &\quad= (2\pi)^{-d_m/2}|\Sigma_m|^{-1/2}c_m^{-d_m} \\ &\qquad\cdot \exp\biggl(-\frac{1}{2c_m^2}(\boldsymbol{\beta}_m-\boldsymbol{\mu}_m)^T\Sigma_m^{-1}(\boldsymbol{\beta}_m-\boldsymbol{\mu}_m)\biggr). \end{aligned} \tag{3}$$

Then,

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

and for suitably large c_m,

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

Hence, as c_m^2 gets larger, $f(m|\mathbf{y})$ 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 tend 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, Smith and Chan (2001), Denison et al. (2002), Hans, Dobra and West (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, Fernández, Ley and Steel (2001) and Liang et al. (2008) for detailed discussion on appropriate choices of 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. Other approaches which have been proposed for specifying default prior distributions under model uncertainty which provide plausible posterior model probabilities include intrinsic priors (Berger and Pericchi, 1996) and, for normal linear models, mixtures of g-priors (Liang et al., 2008). The important effect that any of these prior specifications might have on the parameter posterior distributions within each model has been largely 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. This is a particularly important consideration when posterior or predictive inferences are integrated over models (model-averaging). In such analyses both the prior model probabilities and prior distributions over model parameters can have a significant impact on inferences.

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.

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, Ibrahim and Yiannoutsos (1999) Chen et al. (2003). Within the approach suggested here, prior model probabilities are specified by considering the way in which data yet to be observed might modify one's 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.

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. This allows us to investigate, across a wide class of models, the sensitivity of posterior inferences to the specification of prior model probabilities and prior distributions for model parameters. By using Laplace's method to approximate the posterior marginal likelihood in (4), we obtain, subject to certain regularity conditions (see Kass, Tierney and Kadane, 1988; Schervish, 1995, Section 7.4.3)

$$\begin{aligned} f(m|\mathbf{y}) &\propto f(m)|\Sigma_m|^{-1/2}c_m^{-d_m} f(\mathbf{y}|m, \widehat{\boldsymbol{\beta}}_m) \\ &\quad \cdot \exp\biggl(-\frac{1}{2c_m^2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m)^T \Sigma_m^{-1}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\mu}_m)\biggr) \\ &\quad \cdot |c_m^{-2}\Sigma_m^{-1} - H(\widehat{\boldsymbol{\beta}}_m)|^{-1/2}(1 + O_p(n^{-1})), \end{aligned} \tag{6}$$

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

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

where C is a normalizing constant to ensure that the posterior model probabilities sum to 1 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).

We propose specifying the decomposition of the prior variance matrix $c_m^2\Sigma_m$ so that $|\Sigma_m| = |i(\boldsymbol{\beta}_m)|^{-1}$, resulting in

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

where c_m^{-2} defined as

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

can be interpreted as the number of units of information in the prior.

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) gave 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 $\mathbf{y}$ 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 of 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, Dellaportas and Forster (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

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

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 toward the prior mean to a greater extent than desired. This has a particular impact when model averaging is used to provide predictive inferences (see, e.g., 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. The purpose of decomposing prior model probabilities $f(m)$ in this way is to explicitly specify a direct dependence between these probabilities and the hyperparameters of the prior distributions for the parameters of each model. There is no requirement that $p(m)$ be uniform, and any differences between $f(m)$ for different m which are unrelated to the prior distributions for the model parameters are absorbed in $p(m)$. Often, we might expect $p(m)$ not to depend on the dimensionalities of the models, although we do not prohibit this. With this choice of $f(m)$, (8) becomes

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

where the specification of the base variance Σ_m is not in terms of unit information, the extra term $-\log(|\Sigma_m| \cdot |i(\boldsymbol{\beta}_m)|)/2$ 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 Section 1, we discussed existing approaches for specifying nonuniform $f(m)$ based on considerations such as the desire to control model size. These can easily be incorporated into the specification of nonuniform $p(m)$, if desired. Other possible approaches to specifying or eliciting $p(m)$ are discussed in Sections 4 and 5.

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 $f(m) \propto p(m)|V_m|^{1/2}|i(\boldsymbol{\beta}_m)|^{1/2}$, 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 $\boldsymbol{\beta}_m$. This is not appropriate as a specification for the marginal prior distribution over model space. 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 $m \in M$, $\mathbf{y} \sim N(\mathbf{X}_m\boldsymbol{\beta}_m, \sigma^2 I)$ with the conjugate prior specification

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

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

$$\begin{aligned} &f(m|\mathbf{y}) \\ &\quad \propto f(m)\frac{|V^*|^{1/2}}{|V|^{1/2}} \\ &\qquad \cdot (2\lambda + \mathbf{y}^T\mathbf{y} + \boldsymbol{\mu}^T V^{-1}\boldsymbol{\mu} - \widetilde{\boldsymbol{\beta}}^T (V^*)^{-1}\widetilde{\boldsymbol{\beta}})^{-\alpha-n/2}, \end{aligned}$$

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

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

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

where, with a slight abuse of notation, $i = n^{-1}\mathbf{X}^T\mathbf{X}$ is the unit information matrix multiplied by σ^2. Notice the correspondence between (7) and (16). 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

$$\begin{aligned}\log f(m|\mathbf{y}) &= C - (\alpha + n/2)\log\big(2\lambda + (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \\ &\qquad + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\mu})^T V^{-1}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\mu})\big) \\ &\quad + \log f(m) - \frac{d}{2}\log n - d\log c + O(n^{-1}).\end{aligned} \tag{17}$$

In both (16) and (17) 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 (17). The residual sum of squares is evaluated at the posterior mode, and is penalized 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 $\Sigma^{-1} = i = n^{-1}\mathbf{X}^T\mathbf{X}$, we obtain the prior used by Fernández, Ley and Steel (2001), who also noted the unit information interpretation when $c = 1$ for all m. This is an example of a g-prior (Zellner, 1986).

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) suggested 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

$$f(m) \propto p(m)|V|^{1/2}|i + n^{-1}V^{-1}|^{1/2}, \tag{18}$$

makes (16) 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 (17) implies that model comparison is performed on the basis of BIC.

We note that when the g-prior $\Sigma^{-1} = i = n^{-1}\mathbf{X}^T\mathbf{X}$ is used, together with $\boldsymbol{\mu} = \mathbf{0}$, then the posterior model probability (15) can be written as

$$\begin{aligned}\log f(m|\mathbf{y}) &= C + \log f(m) - \frac{d}{2}\log(n + c^{-2}) - d\log c \\ &\quad - (\alpha + n/2)\log\bigg(2\lambda + \frac{1}{1 + nc^2}\mathbf{y}^T\mathbf{y} \\ &\qquad + \frac{nc^2}{1 + nc^2}S_y^2(1 - R^2)\bigg),\end{aligned} \tag{19}$$

where $S_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ and R^2 is the standard coefficient of determination for the model. For our prior, where $f(m) \propto p(m)c^d$, we obtain

$$\begin{aligned}\log f(m|\mathbf{y}) &= C + \log p(m) - \frac{d}{2}\log(n + c^{-2}) \\ &\quad - (\alpha + n/2)\log\bigg(2\lambda + \frac{1}{1 + nc^2}\mathbf{y}^T\mathbf{y} \\ &\qquad + \frac{nc^2}{1 + nc^2}S_y^2(1 - R^2)\bigg).\end{aligned}$$

The trade-off between model fit, as reflected by R^2, and complexity, measured by d, is immediately apparent, with the complexity penalty tending to BIC as c^{-2} tends to zero. The posterior model probability (19) is similar to expression (5) of Liang et al. (2008). Their approach differs in that they consider the intercept parameter of the linear model separately, giving it an improper uniform prior, as this parameter is common to all models under consideration. Such a specification might also be adopted within our framework, both for linear models and for more general regression models.

4. SPECIFICATION OF $p(m)$ BASED ON 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 $n^{-1/2}$) 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 $O(n^{-1/2})$ neighborhood 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 (nonuniform) specifications for $p(m)$ might be based on matching the posterior model probabilities (8) using prior weights (13) with other information criteria of the form

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

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 since AIC is not consistent), whereas any approach arising from a proper prior must be consistent. Nevertheless, if a penalty function of a particular form is desired for a sample of a specified size n_0, then setting $\log p(m) = \frac{d_m}{2}\{\log n_0 - \psi(n_0)\}$ 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.

Clyde (2000) proposed CIC, a calibrated information criterion, based on a joint specification of (improper) uniform prior distributions for model parameters, together with prior model probabilities

$$f(\boldsymbol{\beta}_m|m) f(m) \propto (2\pi)^{-d_m/2} \left| \frac{n}{c} i(\widehat{\boldsymbol{\beta}}_m) \right|^{1/2},$$

where c is a constant which is determined by constraining the posterior model probabilities to be the same as those which would arise from an alternative information criterion, such as BIC. For our prior, in the limit as $c_m^{-2} \to 0$, we have

$$f(\boldsymbol{\beta}_m|m) f(m) \propto (2\pi)^{-d_m/2} |\Sigma_m|^{-1/2} p(m)$$

so in the case where $|\Sigma_m| = |i(\boldsymbol{\beta}_m)|$ for a value of $\boldsymbol{\beta}_m$ close to the m.l.e. these approaches will yield similar results if $p(m)$ is calibrated to $(n/c)^{d/2}$, which is plausible if $c \propto n$. Note also that, if $p(m) \propto (2\pi)^{d_m/2} |\Sigma_m|^{1/2}$, our prior in this limiting case reduces to a uniform measure over the "parameter space" for $(m, \boldsymbol{\beta}_m)$.

5. ALTERNATIVE ARGUMENTS FOR $f(m) \propto c_m^{d_m}$

The purpose of the following discussion is not 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 behavior of posterior means, as well as by the asymptotic behavior of posterior model probabilities, or associated numerical approximations, as earlier.

5.1 Constant Probability in a Neighborhood of the Prior Mean

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

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

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 (e.g., they imply the same linear predictor in a generalized 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

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

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 neighborhood (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 $p(m) = 2^{d_m/2-1}\Gamma(d_m/2)/\varepsilon^{d_m}$.

5.2 Flattening Prior Densities

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 renormalize. 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 of 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 renormalizing. 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 renormalization 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 $p(m) = (2\pi)^{d_m/2}|i(\boldsymbol{\beta}_m)|^{-1/2}$.

5.3 Bayesian Model Averaging and Shrinkage

Finally, this approach can be justified by considering the behavior 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 β and suppose that $f(y|m_0) = f(y|m_1, \beta_0)$. We assume that the prior for β under m_1 is $N(\beta_0, c^2)$, so the prior mean under model m_1 is the specified value of β under model m_0, and, without loss of generality, we take $\beta_0 = 0$. Under model m_1 the Bayes estimator for β is the posterior mean $E_1(\beta|y)$, which has asymptotic expansion

$$(20) \quad E_1(\beta|y) = \widehat{\beta}\left(1 - \frac{i(\widehat{\beta})}{nc^2}\right) + \frac{a_3}{2i(\widehat{\beta})^2 n} + o(n^{-1}),$$

where na_3 is the third derivative of the log-likelihood, evaluated at $\widehat{\beta}$ (see, e.g., Johnson, 1970; Ghosh, 1994). This illustrates the usual effect of prior variance c^2 and the corresponding prior precision c^{-2} as a shrinkage parameter, with the posterior mean being shrunk away from the m.l.e., with the amount of shrinkage diminishing as $c^{-2} \to 0$. Hence, for fixed y, the posterior mean for β is (asymptotically) monotonic in c^{-2}. Allowing for model uncertainty, we have $E(\beta|y) = f(m_1|y)E_1(\beta|y)$ where

$$(21) \qquad f(m_1|y) = \frac{1}{1 + k(2\pi)^{1/2}cf_1(0|y)},$$

where $f_1(\beta|y)$ is the posterior (marginal) density for β under m_1, and k are the prior odds in favor of m_0 over m_1. Combining (20) and (21), we see that, in $E(\beta|y)$, the model-averaged posterior mean for β, the m.l.e. $\hat{\beta}$ is multiplied by a shrinkage coefficient, $f(m_1|y)E_1(\beta|y)$, which is not a monotonic function of the prior precision for β and hence c^{-2} 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

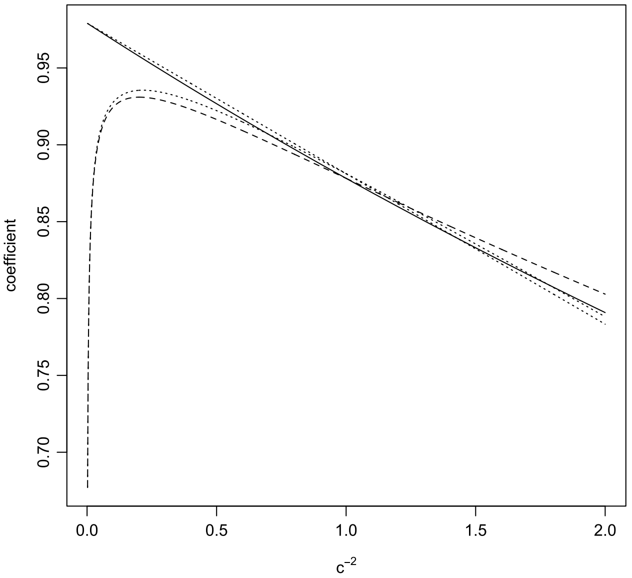


FIG. 1. *Model average coefficient on $\widehat{\beta}$ [evaluated as $\widehat{\beta}/\beta$], for normal likelihood with known error variance, σ^2. The plot here is for $n = 10, \widehat{\beta} = 1, \sigma^2 = 1$. The dashed line is for a uniform prior over models, and the solid line uses prior model probability $f(m_1) \propto c^{-1}$. The dotted lines are approximations based on replacing $(2\pi)^{1/2} f_1(0|y)$ in (21) with its normal approximation $\exp(-\frac{i(\widehat{\beta})n}{2}\widehat{\beta}^2)$, ignoring the dependence, to $O(n^{-1})$, of $f_1(0|y)$ on c^{-2}.*

c^{-2}, for the simple example of a normal distribution with known error variance, and prior odds $k = 1$, corresponding to a uniform prior on model space. Note that a high value of the coefficient on $\widehat{\beta}$ corresponds to low shrinkage. It can be seen that, regardless of the value of c^{-2}, there is a certain amount of shrinkage toward the prior mean and the shrinkage is not a monotone function of c^{-2}. For values of c^{-2} greater than 0.5, the shrinkage to the prior mean is an approximately linearly increasing function of c^{-2} as expected. For small values of c^{-2}, posterior probability is increasingly concentrated on m_0 as c^{-2} decreases (Lindley paradox) and hence the model-averaged estimate is increasingly shrunk to the prior mean. Adopting the approach advocated in this paper has the effect of setting $k \propto c^{-1}$ which mitigates this effect, and returns control over the shrinkage to the analyst.

6. ILLUSTRATED EXAMPLES

We illustrate our approach in a series of simulations and real data applications. For comparison, we also present results under other prior specifications, notably the hyper g-prior of Liang et al. (2008), for which computation is performed using the BAS package; see Clyde (2010).

Section 6.1 illustrates that unit information prior specifications (or other specifications suggesting smaller prior parameter dispersion) can indeed significantly shrink posterior distributions toward zero. This effect suggests that although prior variances based on unit information might have desirable behavior 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.

In Section 6.2 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 for $f(m)$ we adopt our proposed adjusted prior distribution given by (11) with $p(m) = 1$, the prior distribution for the model parameters can be made highly diffuse in a way which does not impact strongly on the posterior model probabilities.

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

6.1 Example 1: A Simple Linear Regression Example

Montgomery, Peck and Vining (2001) investigated 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

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

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 $V_m = c_m^2 n(\mathbf{X}_m^T \mathbf{X}_m)^{-1}$ 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 1 for values of c_m^2 as large as 10^{28}. This behavior provides a source of security with respect to the choice of c_m^2

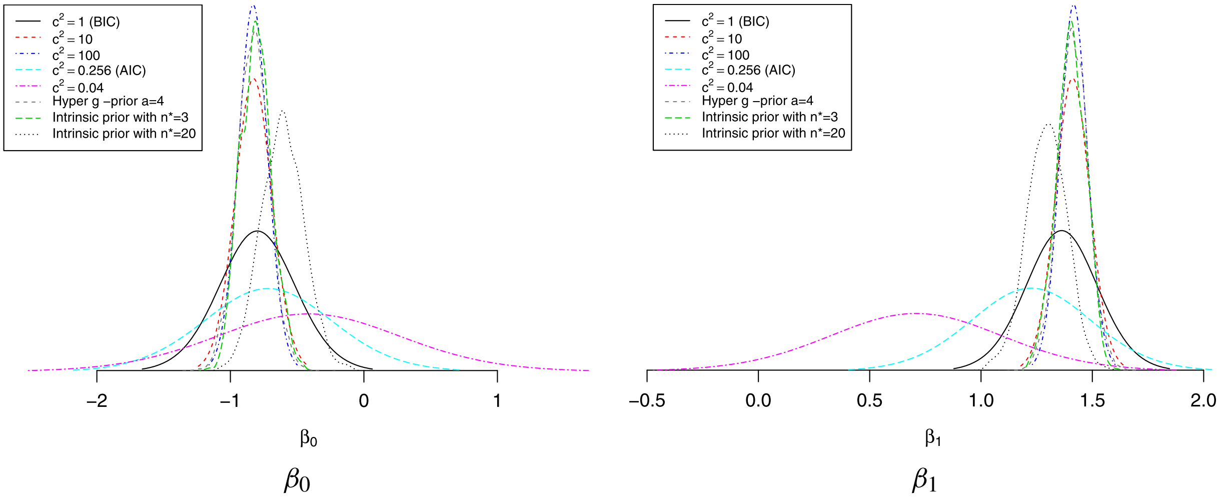


FIG. 2. *Posterior densities of parameters β_0 and β_1 under different prior dispersions*; $c_m^2 = c^2$ *for all models m for Example* 6.1.

and Lindley's paradox, and we use this example to investigate the effect of c_m^2 on the posterior densities of β_0 and β_1; see Figure 2. 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 $c_m^2 = 0.04$; see also George and Foster (2000). 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 was discussed in detail by Paciorek (2006).

We also investigated how the Zellner and Siow (1980) prior and the Liang et al. (2008) hyper g-prior behave in this example. With the recommended hyperparameter values $2 < a \leq 4$, these priors produced posterior densities close to the low information g-prior with $c_m^2 = 100$; see Figure 2. The results are quite robust across this range for a and, for example, quite large values of a, around 20, are required before the level of shrinkage becomes comparable to the unit information g-prior. Hence inferences arising from the hyper-g prior are quite robust across the recommended range of hyperparameter values.

Finally, we examined the effect of intrinsic priors on posterior distributions for model parameters. We adopted the approach of Perez and Berger (2002) to construct an intrinsic (or expected posterior) prior by setting as a baseline prior the g-prior with $c^2 = 100$ and the null model as a reference. For this simple linear regression model the minimal training sample has size $n^* = 3$. The resulting posterior distributions of β_0 and β_1, also shown in Figure 2, are in close agreement with the baseline g-prior. However, in variable selection problems the minimal training sample is usually set so that the full model can be estimated. Hence, the value of n^* could be much higher if more covariates were available and this would affect the prior variance of the parameters. As an example, we have calculated the posterior densities of β_0 and β_1 when $n^* = 20$, also displayed in Figure 2. The effect of the prior densities to the posterior distributions is dramatic. This nicely illustrates the effect of the training sample size in intrinsic priors; see the relevant discussion in Berger and Pericchi (2004).

We now investigate the effect of prior specification when prediction is of primary interest. A common way of evaluating predictive performance is to compute the negative cross-validation score (see Geisser and Eddy, 1979) given by

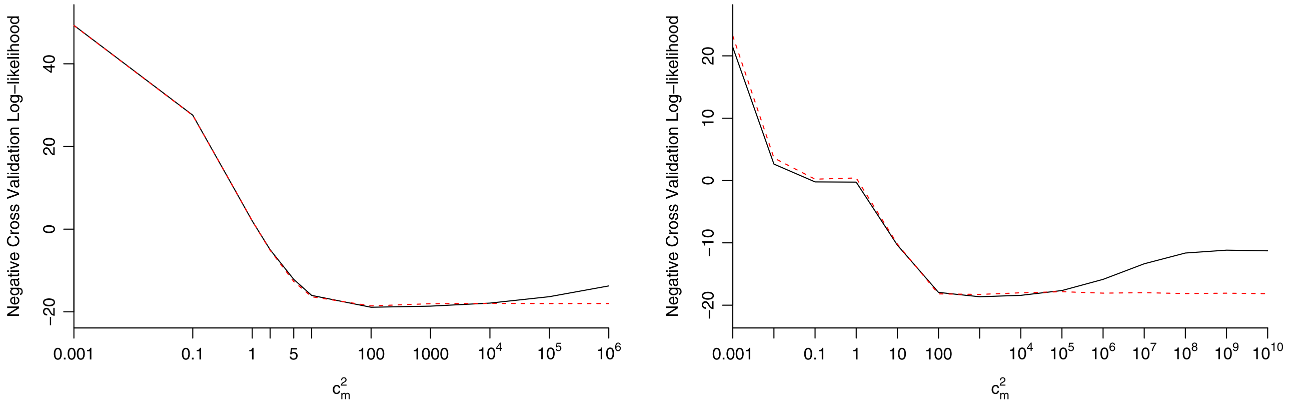
$$S = -\sum_{j=1}^{n} \log f^p(j),$$

where

$$f^p(j) = \sum_{m \in M} f(m) f(y_j | \mathbf{y}_{\setminus j}, m)$$

is the model-averaged predictive density of observation y_j given the rest of the data $\mathbf{y}_{\setminus j}$. Lower values of S indicate greater predictive accuracy. Following Gelfand (1996) we estimate $f^p(j)$ from an MCMC sample by the inverse of the posterior (over $m, \boldsymbol{\beta}_m$) mean of the inverse predictive density of observation j.

We generated three additional covariates that have correlation coefficients 0.99, 0.97 and 0.89 with x



(a) g-prior ($V_m = c_m^2 n(\mathbf{X}_m^T\mathbf{X}_m)^{-1}$).

(b) Independence prior ($V_m = c_m^2 \mathbf{I}_{d_m}$).

FIG. 3. *Negative cross-validation log-likelihood for two prior dispersion structures with uniform prior* (*solid line*) *and adjusted prior* (*dashed line*) *for Example* 6.1.

and performed the same model determination exercise. Posterior model probabilities for all models were calculated for all models under consideration. We used a g-prior with $V_m = c_m^2 n(\mathbf{X}_m^T\mathbf{X}_m)^{-1}$ and an independent prior with $V_m = c_m^2\mathbf{I}_{d_m}$. For the uniform prior on models combined with the unit information prior obtained by $c_m^2 = 1$, S is far away from the minimum value achieved for higher values of c_m^2; see Figure 3(a). For $c_m^2 > 10^5$, S increases due to the effect of Lindley's paradox focusing posterior probability on models that are unrealistically simple. On the other hand, our proposed adjusted prior specification achieves the maximum predictive ability for any large value of c_m^2; see Figure 3(b). The same exercise was also repeated for the hyper-g prior for various values of the hyperparameter a. The corresponding negative cross-validation score was close to the stabilized value of the g-prior and it was proven to be very robust for a wide range of values of a. Only for a very close to 2, did predictive ability start to deteriorate in a similar fashion to the g-prior.

This simulated data exercise does indicate that predictive ability can be optimized if highly dispersed prior parameter densities are chosen together with the adjusted prior over model space. Alternatively, in this example, the hyper-g family is sufficiently robust to simultaneously provide a diffuse prior for model parameters, together with reasonable behavior under model uncertainty.

6.2 Example 2: Simulated Regressions

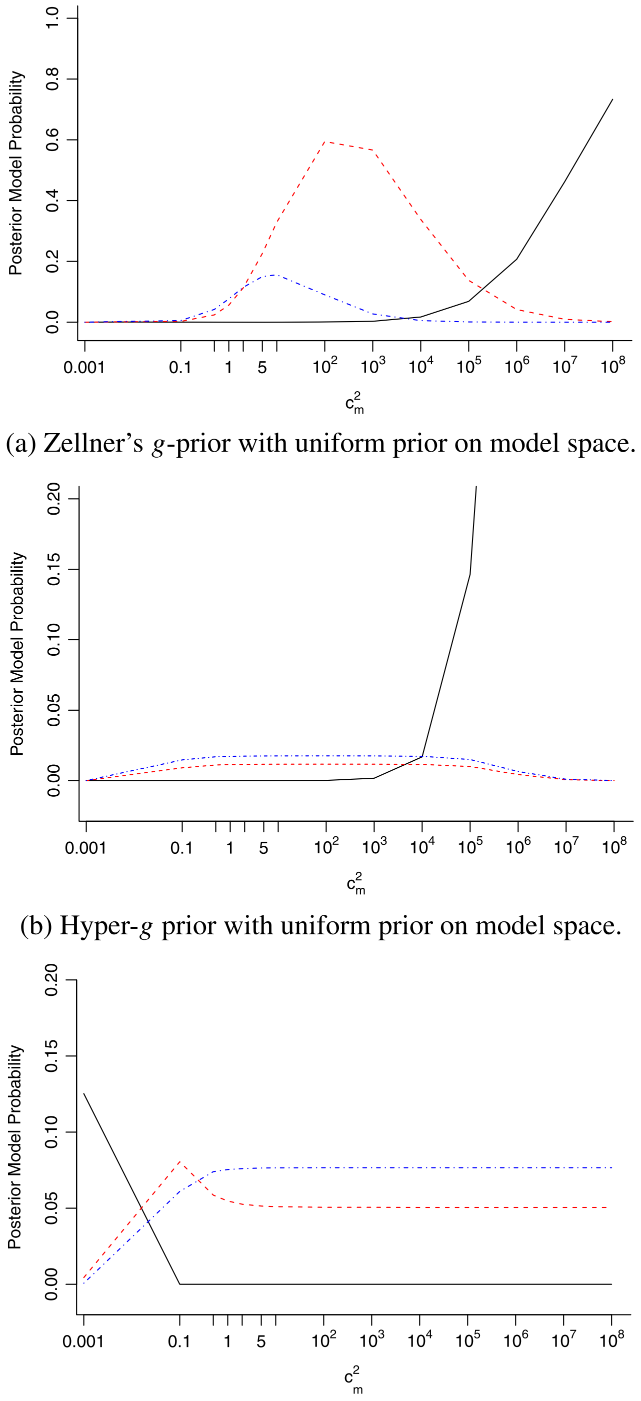
We now consider the first simulated dataset of Dellaportas, Forster and Ntzoufras (2002) based on $n = 50$ observations of 15 standardized independent normal covariates X_j, $j = 1, \dots, 15$, and a response variable Y generated as

$$Y \sim N(X_4 + X_5,\ 2.5^2). \tag{22}$$

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 behavior is exhibited either when Σ_m is specified as $\Sigma_m = n(\mathbf{X}_m^T\mathbf{X}_m)^{-1}$ (described below) or as $\Sigma_m = \mathbf{I}_{d_m}$.

Figure 4(a) and (b), illustrates the behavior of the posterior model probabilities, under a uniform prior on model space, of three indicative models. For the parameters we used the g-prior and the hyper-g prior with $c_m^2 = 2n^{-1}/(a-2)$ obtained by equating the shrinkage proportion $g/(g-1)$ of the g-prior with its prior mean under the hyper-g prior. The effect of Lindley's paradox is more evident for the g-prior where all posterior probabilities are quite sensitive to the values of c_m^2 while the hyper-g prior demonstrates a remarkable robustness for a wide range of prior parameter values and only for quite large values of c_m^2 which correspond to values of a close to 2 is Lindley's paradox exhibited. We note that the hyper-g prior seems to result in increased uncertainty on model space resulting in lower posterior model probabilities for the higher posterior probability models.

By contrast, using the adjusted prior in Figure 4(c) identifies $1 + X_4 + X_5 + X_{12}$ as the highest probability model for any value of $c_m^2 > 1$. Note that, when $\Sigma_m = n(\mathbf{X}_m^T\mathbf{X}_m)^{-1}$, $c_m^2 = 1$ represents the dispersion induced by the unit information prior. Similarly, Figure 5 summarizes the posterior inclusion probability of each variable X_j. Again, for the uniform prior these

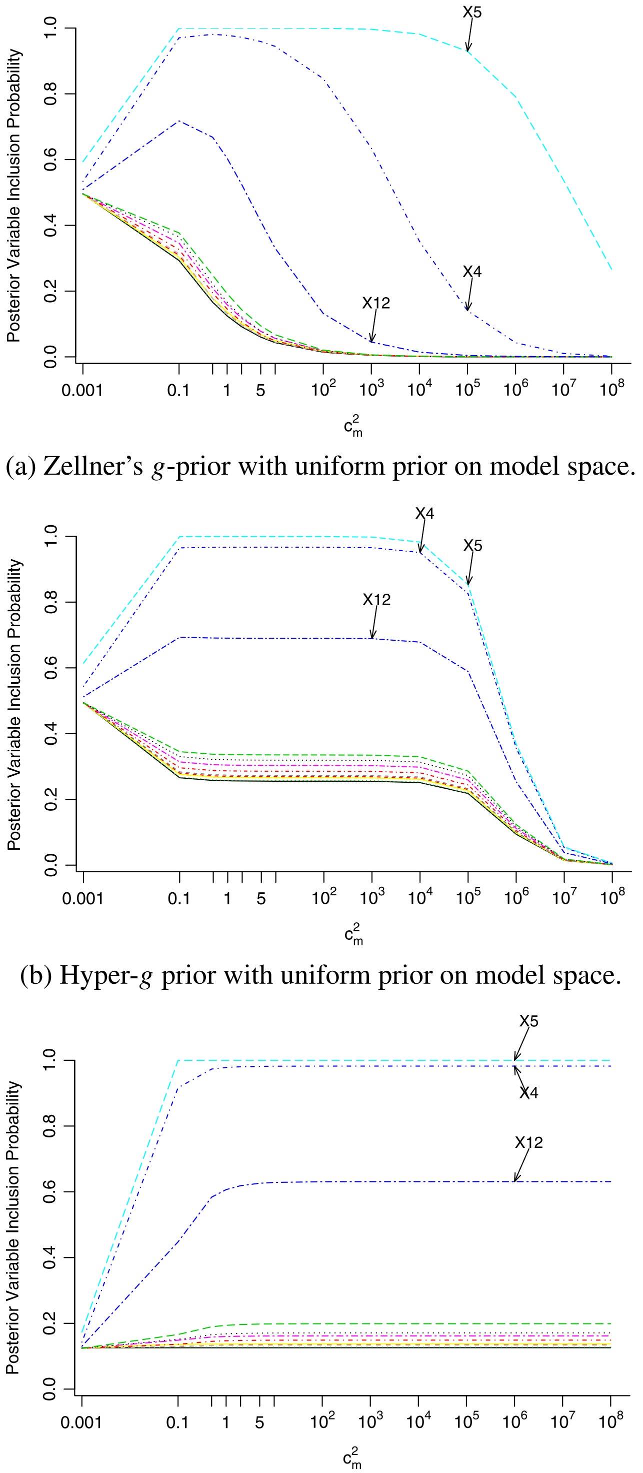


(a) Zellner's g-prior with uniform prior on model space.

(b) Hyper-g prior with uniform prior on model space.

(c) Zellner's g-prior with adjusted prior on model space.

FIG. 4. *Posterior model probabilities under different prior dispersions for the Dellaportas, Forster and Ntzoufras (2002) dataset of Section 6.2 generated using (22). 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) Zellner's g-prior with uniform prior on model space.

(b) Hyper-g prior with uniform prior on model space.

(c) Zellner's g-prior with adjusted prior on model space.

FIG. 5. *Posterior variable inclusion probabilities under different prior dispersions for the Dellaportas, Forster and Ntzoufras (2002) dataset of Section 6.2 generated using (22).*

probabilities are sensitive to changes in c_m^2 across its range, whereas the adjusted prior produces stable results for $c_m^2 > 1$.

In a more detailed simulation study, we repeated the above analysis by generating 100 datasets of the same model. The distribution of the posterior model probabilities over the 100 simulated datasets reinforces the findings of the one-sample based simulation. We also repeated the above simulation experiment with a more challenging simulated dataset based on a simulation structure suggested by Nott and Kohn (2005). Each dataset consisted of $n = 50$ observations and $p = 15$ covariates and one response generated using the following sampling scheme:

$$(23)\qquad \left\{\begin{array}{l} X_j \sim N(0,1) \quad \text{for } j = 1,\dots,10 \\ X_j \sim N(0.3X_1 + 0.5X_2 + 0.7X_3 \\ \qquad + 0.9X_4 + 1.1X_5, 1) \\ \quad \text{for } j = 11,\dots,15 \\ \quad Y \sim N(4 + 2X_1 - X_5 + 1.5X_7 \\ \qquad + X_{11} + 0.5X_{13}, 2.5^2) \end{array}\right\}.$$

The general conclusions of this study are in close agreement with the results obtained above. Further details are available in the electronic supplement which is available at http://stat-athens.aueb.gr/~jbn/papers/paper24.htm.

6.3 Example 3: A 3 x 2 x 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):

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

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^T$, $\mathbf{X}$ is the $n \times n$ design matrix of the full model, $\boldsymbol{\beta} = (\boldsymbol{\beta}_j; j \in \mathcal{V})$ is an $n \times 1$ 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

$$(24)\qquad \boldsymbol{\beta}_j \sim N\left(\boldsymbol{\mu}_j, \quad k_j^2(\mathbf{X}_j^T\mathbf{X}_j)^{-1}\right)$$

with $\boldsymbol{\mu}_j$ being a vector of zeros and $k_j^2 = 2n$ for all $j \in \mathcal{V} = \{\varnothing, O, H, A, OH, OA, HA, OHA\}$; 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 $\mathcal{V} = \{\varnothing, O, H, A, OH, HA\}$ should be considered. Moreover, the term $\boldsymbol{\beta}_{HA}$ is nonzero 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 (24) 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 $j \in \{\varnothing, O, H, A, OH\}$. 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 $j \in \{\varnothing, O, H, A, OH\}$. 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 adjusted prior we utilize the prior mean $\boldsymbol{\mu}_m$ of $\boldsymbol{\beta}_m$ resulting in

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

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.

Posterior model probabilities (estimated using reversible jump MCMC) for all prior specifications are presented in Table 1. The top right panel of the table illustrates the striking effect of informative parameter priors on posterior model probabilities. The difficulty of making joint inferences on parameter and model space is evident by inspecting the sensitivity of model probabilities to different priors. However, the specification for adjusting the prior model probabilities has the effect 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,

TABLE 1

Prior and posterior model probabilities under different parameter and model prior densities for Example 6.3

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

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

BARTLETT, M. S. (1957). Comment on D. V. Lindley's statistical paradox. *Biometrika* **44** 533–534.

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065

BERGER, J. O. and PERICCHI, L. R. (2004). Training samples in objective Bayesian model selection. *Ann. Statist.* **32** 841–869. MR2065191

CHEN, M.-H., IBRAHIM, J. G. and YIANNOUTSOS, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 223–242. MR1664057

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. *J. Statist. Plann. Inference* **111** 57–76. MR1955872

CHIPMAN, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.* **24** 17–36. MR1394738

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 65–134. IMS, Beachwood, OH. MR2000752

CLYDE, M. (2000). Model uncertainty and health effect studies for particulate matter. *Environmetrics* **11** 745–763.

CLYDE, M. (2010). The BAS Package: Bayesian model averaging and stochastic search using Bayesian adaptive sampling (Version 0.91). Available at http://www.stat.duke.edu/~clyde/BAS/.

CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference* **138** 888–900. MR2416869

DAWID, A. P. (2011). Posterior model probabilities. In *Philosophy of Statistics* (P. S. Bandyopadhyay and M. Forster, eds.) 607–630. Elsevier, New York.

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. MR1723782

DELLAPORTAS, P., FORSTER, J. J. and NTZOUFRAS, I. (2002). On Bayesian model and variable selection using MCMC. *Stat. Comput.* **12** 27–36.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. and SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, Chichester. MR1962778

FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. MR0529531

GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.) 145–161. Chapman & Hall, London. MR1397969

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

GHOSH, J. K. (1994). *Higher Order Asymptotics*. IMS, Hayward, CA.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GREEN, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. In *Highly Structured Stochastic Systems*. *Oxford Statist. Sci. Ser.* **27** 179–206. Oxford Univ. Press, Oxford. MR2082410

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc.* **102** 507–516. MR2370849

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, New York. MR0187257

JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41** 851–864. MR0263198

KADANE, J. B. and LAZAR, N. A. (2004). Methods and criteria for model selection. *J. Amer. Statist. Assoc.* **99** 279–290. MR2061890

KASS, R. E., TIERNEY, L. and KADANE, J. B. (1988). Asymptotics in Bayesian computation. In *Bayesian Statistics*, 3 (*Valencia*, 1987). *Oxford Sci. Publ.* 261–278. Oxford Univ. Press, New York. MR1008051

KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934. MR1354008

KNUIMAN, M. W. and SPEED, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics* **44** 1061–1071. MR0981000

KOHN, R., SMITH, M. and CHAN, D. (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput.* **11** 313–322. MR1863502

LAUD, P. W. and IBRAHIM, J. G. (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B* **57** 247–262. MR1325389

LAUD, P. W. and IBRAHIM, J. G. (1996). Predictive specification of prior model probabilities in variable selection. *Biometrika* **83** 267–274. MR1439783

LEY, E. and STEEL, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *J. Appl. Econometrics* **24** 651–674. MR2675199

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

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.) 91–100. Springer, Berlin.

MONTGOMERY, D. C., PECK, E. A. and VINING, G. G. (2001). *Introduction to Linear Regression Analysis*, 3rd ed. Wiley, New York. MR1820113

NOTT, D. J. and KOHN, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92** 747–763. MR2234183

NTZOUFRAS, I., DELLAPORTAS, P. and FORSTER, J. J. (2003). Bayesian variable and link determination for generalised linear models. *J. Statist. Plann. Inference* **111** 165–180. MR1955879

PACIOREK, C. J. (2006). Misinformation in the conjugate prior for the linear model with implications for free-knot spline modelling. *Bayesian Anal.* **1** 375–383 (electronic). MR2221270

PÉREZ, J. M. and BERGER, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* **89** 491–511. MR1929158

PERICCHI, L. R. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models. *Biometrika* **71** 575–586. MR0775404

POSKITT, D. S. and TREMAYNE, A. R. (1983). On the posterior odds of time series models. *Biometrika* **70** 157–162. MR0742985

RAFTERY, A. E. (1995). Bayesian model selection for social research (with discussion). In *Sociological Methodology* 1995 (P. V. Marsden, ed.) 111–196. Blackwell, Oxford.

RAFTERY, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83** 251–266. MR1439782

ROBERT, C. P. (1993). A note on Jeffreys–Lindley paradox. *Statist. Sinica* **3** 601–608. MR1243404

SCHERVISH, M. J. (1995). *Theory of Statistics*, 2nd ed. Springer, New York. MR1354146

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

VOLINSKY, C. T. and RAFTERY, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56** 256–262.

WILSON, M. A., IVERSEN, E. S., CLYDE, M. A., SCHMIDLER, S. C. and SCHILDKRAUT, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *Ann. Appl. Stat.* **4** 1342–1364. MR2758331

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*. *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

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