ELSEVIER

Journal of Econometrics 100 (2001) 381–427

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

Benchmark priors for Bayesian model averaging

Carmen Fernández[a], Eduardo Ley[b,*], Mark F.J. Steel[c]

[a]*School of Mathematics and Statistics, University of St. Andrews, St. Andrews KY16 9SS, UK*
[b]*International Monetary Fund, 700 19th St NW, Washington, DC 20431, USA*
[c]*Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury CT2 7NF, UK*

Received 13 April 1999; received in revised form 13 July 2000; accepted 1 August 2000

Abstract

In contrast to a posterior analysis given a particular sampling model, posterior model probabilities in the context of model uncertainty are typically rather sensitive to the specification of the prior. In particular, 'diffuse' priors on model-specific parameters can lead to quite unexpected consequences. Here we focus on the practically relevant situation where we need to entertain a (large) number of sampling models and we have (or wish to use) little or no subjective prior information. We aim at providing an 'automatic' or 'benchmark' prior structure that can be used in such cases. We focus on the normal linear regression model with uncertainty in the choice of regressors. We propose a partly non-informative prior structure related to a natural conjugate g-prior specification, where the amount of subjective information requested from the user is limited to the choice of a single scalar hyperparameter g_{0j}. The consequences of different choices for g_{0j} are examined. We investigate theoretical properties, such as consistency of the implied Bayesian procedure. Links with classical information criteria are provided. More importantly, we examine the finite sample implications of several choices of g_{0j} in a simulation study. The use of the MC^3 algorithm of Madigan and York (Int. Stat. Rev. 63 (1995) 215), combined with efficient coding in Fortran, makes it feasible to conduct large simulations. In addition to posterior criteria, we shall also compare the predictive performance of different priors. A classic example concerning the economics of crime will also be provided and contrasted with results in the literature. The main findings of the

* Corresponding author. Tel.: + 1-202-623-6107; fax: + 1-202-589-6107.
E-mail address: eley@imf.org (E. Ley).

0304-4076/01/$ - see front matter
PII: S0304-4076(00)00076-2

paper will lead us to propose a 'benchmark' prior specification in a linear regression context with model uncertainty.

JEL classification: C11; C15

Keywords: Bayes factors; Markov chain Monte Carlo; Posterior odds; Prior elicitation

1. Introduction

The issue of model uncertainty has permeated the econometrics and statistics literature for decades. An enormous volume of references can be cited (only a fraction of which is mentioned in this paper), and special issues of the *Journal of Econometrics* (1981, Vol. 16, No. 1) and *Statistica Sinica* (1997, Vol. 7, No. 2) are merely two examples of the amount of interest this topic has generated in the literature. From a Bayesian perspective, dealing with model uncertainty is conceptually straightforward: the model is treated as a further parameter which lies in the set of models entertained (the model space). A prior now needs to be specified for the parameters within each model as well as for the models themselves, and Bayesian inference can be conducted in the usual way, with one level (the prior on the model space) added to the hierarchy — see, e.g., Draper (1995) and the ensueing discussion. Unfortunately, the influence of the prior distribution, which is often straightforward to assess for inference given the model, is much harder to identify for posterior model probabilities. It is acknowledged — e.g., Kass and Raftery (1995), George (1999) — that posterior model probabilities can be quite sensitive to the specification of the prior distribution.

In this paper, we consider a particular instance of model uncertainty, namely uncertainty about which variables should be included in a linear regression problem with k available regressors. A model here will be identified by the set of regressors that it includes and, thus, the model space consists of 2^k elements.[1] Given the issue of sensitivity to the prior distribution alluded to above, the choice of prior is quite delicate, especially in the absence of substantial prior knowledge. Our aim here is to come up with a prior distribution that leads to sensible results, in the sense that data information dominates prior assumptions. Whereas we acknowledge the merits of using substantive prior information whenever available, we shall be concerned with providing the applied researcher with a 'benchmark' method for conducting inference in situations where incorporating such information into the analysis is deemed impossible, impractical or

[1] Of course, more models arise if we consider other aspects of model specification, but this will not be addressed here. See, e.g., Hoeting et al. (1995, 1996) for treatments of variable transformations and outliers, respectively.

undesired. In addition, this provides a useful backdrop against which results arising from Bayesian analyses with informative priors could be contrasted.

We will focus on Bayesian model averaging (BMA), rather than on selecting a single model. BMA follows directly from the application of Bayes' theorem in the hierarchical model described in the first paragraph, which implies mixing over models using the posterior model probabilities as weights. This is very reasonable as it allows for propagation of model uncertainty into the posterior distribution and leads to more sensible uncertainty bands. From a decision-theory point of view, Min and Zellner (1993) show that such mixing over models minimizes expected predictive squared error loss, provided the set of models under consideration is exhaustive. Raftery et al. (1997) state that BMA is optimal if predictive ability is measured by a logarithmic scoring rule. The latter result also follows from Bernardo (1979), who shows that the usual posterior distribution leads to maximal expected utility under a logarithmic proper utility function. Such a utility function was argued by Bernardo (1979) to be 'often the more appropriate description for the preferences of a scientist facing an inference problem'. Thus, in the context of model uncertainty, the use of BMA follows from sensible utility considerations. This is the scenario that we will focus on. However, our results should also be useful under other utility structures that lead to decisions different from model averaging — e.g. model selection. This is because the posterior model probabilities will intervene in the evaluation of posterior expected utility. Thus, finding a prior distribution that leads to sensible results in the absence of substantive prior information is relevant in either setting.

Broadly speaking, we can distinguish three strands of related literature in the context of model uncertainty. Firstly, we mention the fundamentally oriented statistics and econometrics literature on prior elicitation and model selection or model averaging, such as exemplified in Box (1980), Zellner and Siow (1980), Draper (1995) and Phillips (1995) and the discussions of these papers. Secondly, there is the recent statistics literature on computational aspects. Markov chain Monte Carlo methods are proposed in George and McCulloch (1993), Madigan and York (1995), Geweke (1996) and Raftery et al. (1997), while Laplace approximations are found in Gelfand and Dey (1994) and Raftery (1996). Finally, there exists a large literature on information criteria, often in the context of time series, see, e.g., Hannan and Quinn (1979), Akaike (1981), Atkinson (1981), Chow (1981) and Foster and George (1994). This paper provides a unifying framework in which these three areas of research will be discussed.

In line with the bulk of the literature, the context of this paper will be normal linear regression with uncertainty in the choice of regressors. We abstract from any other issue of model specification. We present a prior structure that can reasonably be used in cases where we have (or wish to use) little prior information, partly based on improper priors for parameters that are common to all models, and partly on a g-prior structure as in Zellner (1986). The prior is not in

the natural-conjugate class, but is such that marginal likelihoods can still be computed analytically. This allows for a simple treatment of potentially very large model spaces through Markov chain Monte Carlo model composition (MC^3) as introduced in Madigan and York (1995). In contrast to some of the priors proposed in the literature, the prior we propose leads to valid conditioning in the posterior distribution (i.e., the latter can be interpreted as a conditional distribution given the observables) as it avoids dependence on the values of the response variable. The only hyperparameter left to elicit in our prior is a scalar g_{0j} for each of the models considered. Theoretical properties, such as consistency of posterior model probabilities, are linked to functional dependencies of g_{0j} on sample size and the number of regressors in the corresponding model. In addition (and perhaps more importantly), we conduct an empirical investigation through simulation. This will allow us to suggest specific choices for g_{0j} to the applied user. As we have conducted a large simulation study, efficient coding was required. This code (in Fortran-77) has been made publicly available on the World Wide Web.[2]

Section 2 introduces the Bayesian model and the practice of Bayesian model averaging. The prior structure is explained in detail in Section 3, where expressions for Bayes factors are also given. The setup of the empirical simulation experiment is described in Section 4, while results are provided in Section 5. Section 6 presents an illustrative example using the economic model of crime from Ehrlich (1973, 1975), and Section 7 gives some concluding remarks and practical recommendations. The appendix presents results about asymptotic behaviour of Bayes factors.

2. The model and Bayesian model averaging

We consider n independent replications from a linear regression model with an intercept, say α, and k possible regression coefficients grouped in a k-dimensional vector β. We denote by Z the corresponding $n \times k$ design matrix and we assume that $r(\iota_n : Z) = k + 1$, where $r(\cdot)$ indicates the rank of a matrix and ι_n is an n-dimensional vector of 1's.

This gives rise to 2^k possible sampling models, depending on whether we include or exclude each of the regressors. In line with the bulk of the literature in this area — see, e.g., Mitchell and Beauchamp (1988) George and McCulloch

[2] Our programs, which can be found at mcmcmc.freeyellow.com should be slightly adapted before they can be used in other problems. More flexible software to implement the approach in Smith and Kohn (1996) can be found at www.agsm.unsw.edu.au/ ~ mikes/software.html, whereas the BMA webpage of Chris Volinsky at www.research.att.com/ ~ volinsky/bma.html lists various resources of relevance to BMA.

(1993) and Raftery et al. (1997) – exclusion of a regressor means that the corresponding element of β is zero. Thus, a model M_j, $j = 1, \ldots, 2^k$, contains $0 \leqslant k_j \leqslant k$ regressors and is defined by

$$y = \alpha \iota_n + Z_j \beta_j + \sigma \varepsilon, \tag{1.1}$$

where $y \in \Re^n$ is the vector of observations. In (1.1), Z_j denotes the $n \times k_j$ submatrix of Z of relevant regressors, $\beta_j \in \Re^{k_j}$ groups the corresponding regression coefficients and $\sigma \in \Re_+$ is a scale parameter. Furthermore, we shall assume that ε follows an n-dimensional normal distribution with zero mean and identity covariance matrix.

We now need to specify a prior distribution for the parameters in (1.1). This distribution will be given through a density function

$$p(\alpha, \beta_j, \sigma | M_j). \tag{1.2}$$

In Section 2, we shall consider specific choices for the density in (1.2) and examine the resulting Bayes factors. We group the zero components of β under M_j in a vector $\beta_{\sim j} \in \Re^{k-k_j}$, i.e.

$$P_{\beta_{\sim j} | \alpha, \beta_j, \sigma, M_j} = P_{\beta_{\sim j} | M_j} = \text{Dirac at } (0, \ldots, 0). \tag{1.3}$$

We denote the space of all 2^k possible models by $\mathcal{M}$, thus

$$\mathcal{M} = \{M_j : j = 1, \ldots, 2^k\}. \tag{1.4}$$

In a Bayesian framework, dealing with model uncertainty is, theoretically, perfectly straightforward: we simply need to put a prior distribution over the model space $\mathcal{M}$

$$P(M_j) = p_j, \quad j = 1, \ldots, 2^k, \quad \text{with } p_j > 0 \text{ and } \sum_{j=1}^{2^k} p_j = 1. \tag{1.5}$$

Thus, we can think of the model in (1.1)–(1.5) as the usual linear regression model where all possible regressors are included, but where the prior on β has a mixed structure, with a continuous part and a discrete point mass at zero for each element. In other words, the model index M_j really indicates that certain elements of β (namely $\beta_{\sim j}$) are set to zero, and, as discussed in Poirier (1985), we always condition on the full set of available regressors.

With this setup, the posterior distribution of any quantity of interest, say Δ, is a mixture of the posterior distributions of that quantity under each of the models with mixing probabilities given by the posterior model probabilities. Thus,

$$P_{\Delta | y} = \sum_{j=1}^{2^k} P_{\Delta | y, M_j} P(M_j | y), \tag{1.6}$$

provided Δ has a common interpretation across models. This procedure, which is typically referred to as Bayesian model averaging (BMA), is in fact the standard Bayesian solution under model uncertainty, since it follows from direct application of Bayes' theorem to the model in (1.1)–(1.5) — see, e.g., Leamer (1978), Min and Zellner (1993), Osiewalski and Steel (1993) and Raftery et al. (1997).

Posterior model probabilities are given by

$$P(M_j|y) = \frac{l_y(M_j)P(M_j)}{\sum_{h=1}^{2^k} l_y(M_h)P(M_h)} = \left(\sum_{h=1}^{2^k} \frac{P(M_h)}{P(M_j)} \frac{l_y(M_h)}{l_y(M_j)} \right)^{-1}, \tag{1.7}$$

where $l_y(M_j)$, the marginal likelihood of model M_j, is obtained as

$$l_y(M_j) = \int p(y|\alpha, \beta_j, \sigma, M_j) p(\alpha, \beta_j, \sigma|M_j)\, \mathrm{d}\alpha\, \mathrm{d}\beta_j\, \mathrm{d}\sigma \tag{1.8}$$

with $p(y|\alpha, \beta_j, \sigma, M_j)$ and $p(\alpha, \beta_j, \sigma|M_j)$ defined through (1.1) and (1.2), respectively.

Two difficult questions here are how to compute $P(M_j|y)$ and how to assess the influence of our prior assumptions on the latter quantity. Substantial research effort has gone into examining each of them:

In cases where $l_y(M_j)$ can be derived analytically, the computation of $P(M_j|y)$ is, in theory, straightforward, through direct application of (1.7). However, the large number of terms (2^k) involved in the latter expression often makes this computation practically infeasible. A common approach is to resort to an MCMC algorithm, by which we generate draws from a Markov chain on the model space $\mathscr{M}$ with the posterior model distribution as its stationary distribution. An estimate of (1.7) is then constructed on the basis of the models visited by the chain. An important example of this is the MC^3 methodology of Madigan and York (1995), which uses a Metropolis-Hastings updating scheme — see, e.g., Chib and Greenberg (1995). MC^3 was implemented in the context of BMA in linear regression models by Raftery et al. (1997), who consider a natural conjugate prior structure in (1.2). The latter paper also proposes an application of the Occam's window algorithm of Madigan and Raftery (1994) for deterministically finding the models whose posterior probability is above a certain threshold. Under a g-prior distribution for the regression coefficients, the use of the fast updating scheme of Smith and Kohn (1996) in combination with the Gray code order, allows for exhaustive evaluation of all 2^k terms in (1.7) when k is less than about 25 — see George and McCulloch (1997).

Computing $P(M_j|y)$ is a more complex problem when analytical evaluation of $l_y(M_j)$ is not available. In that case, the reversible jump methodology of Green (1995), which extends usual Metropolis-Hastings methods to spaces of variable dimension, could be applied to construct a Markov chain jointly over parameter

and model space. An alternative approach was proposed by George and McCulloch (1993), who instead of zero restrictions in (1.3), assume a continuous prior distribution concentrated around zero for these coefficients. In this way, they get around the problem of a parameter space of varying dimension and are still able to propose a Gibbs sampling algorithm to generate a Markov chain. Their approach is based on a zero-mean normal prior for β given M_j, where large and small variances are, respectively, allocated to regression coefficients included in and 'excluded' from M_j. Thus, they are required to choose two prior variances, and results are typically quite sensitive to this choice. As the ratio of the variances becomes large, the mixing of the chain will often be quite slow. An alternative Gibbs sampler that can deal with prior point mass at zero and displays better mixing behaviour was proposed in Geweke (1996). A deterministic approach in the vein of Occam's window was taken by Volinsky et al. (1997), who approximate the value of $l_y(M_j)$ and use a modified leaps-and-bounds algorithm to find the set of models to average over (i.e., the models with highest posterior probability).

Apart from purely computational aspects, just described, the issue of choosing a 'sensible' prior distribution seems further from being resolved. From (1.7) it is clear that the value of $P(M_j|y)$ is determined by the prior odds $[P(M_h)/P(M_j)]$ and the Bayes factors $[B_{hj} \equiv l_y(M_h)/l_y(M_j)]$ of each of the entertained models versus M_j. Bayes factors are known to be rather sensitive to the choice of the prior distributions for the parameters within each model. Even asymptotically, the influence of this distribution does not vanish — see, e.g., Kass and Raftery (1995) and George (1999). Thus, under little (or absence of) prior information, the choice of the distribution in (1.2) is a very thorny question. Furthermore, the usual recourse to improper 'non-informative' priors does not work in this situation, since improper priors cannot be used for model-specific parameters [attempts to overcome this include the explicit or implicit use of training samples, using, e.g., intrinsic Bayes factors as in Berger and Pericchi (1996) or fractional Bayes factors as in O'Hagan (1995), which, although conceptually quite interesting, suffer from a number of inconsistencies]. As a consequence, most of the literature has focussed on 'weakly informative' proper priors, which are often data-dependent through the response variable — as the prior in, e.g., Raftery et al. (1997). George and Foster (1997) propose an empirical Bayes approach (in a case with known σ) to elicit prior hyperparameters, in order to avoid the computational difficulties of a full Bayesian analysis with a further level of hierarchy. Whilst we do not wish to detract from the potential usefulness of data-dependent priors in certain contexts, we note that they do not allow for valid conditioning, in the sense that the posterior distribution cannot be interpreted as a conditional distribution given the observables (although the hope, of course, is that the product of likelihood and 'prior' still constitutes a suitable basis for inference in such cases). Here, we focus on priors that avoid dependence on the values of the response variable and, thus, avoid this (in our view,

undesirable) property. We will propose certain priors and study their behaviour in comparison with other priors previously considered in the literature.

As a final remark before concluding this section, we note that, in line with the majority of recent Bayesian literature in this area, we consider a prior distribution that allows for the actual exclusion of regressors from some of the models — see (1.3). For us, the rationale behind this choice is that, when faced with a modelling scenario with uncertainty in the choice of covariates, the researcher will often ask herself questions of the form 'Does the exclusion of certain subsets of regressors lead to a sensible model?', thus interpreting the exclusion of regressors not like a dogmatic belief that such regressors have no influence whatsoever on the outcome of the process being modelled but, rather, as capturing the idea that the model which excludes those regressors is a sensible one. Just how sensible a model is will be quantified by its posterior probability, which combines prior information (or lack of it) with data information via Bayes' theorem. Of course, there might be situations in which utility considerations — e.g., cost of collecting regressors versus their predictive ability, or some other consideration specific to that particular problem — dictate that certain regressors be dropped from the model even if their inclusion is sensible by the criterion mentioned above. In such cases, the use of a continuous prior concentrated around zero — as in George and McCulloch (1993) — instead of (1.3) or, as a Referee suggested, conducting continuous inference about the full vector β followed by removal of regressors according to utility considerations, could be preferable. However, this paper will not consider design issues and, as mentioned in the Introduction, focusses on the case where neither substantive prior information nor a problem-specific decision theory framework are available, rendering our approach more natural. For more comments on the issue of discrete versus continuous priors, see Raftery et al. (1996) and the ensueing discussion.

3. Priors for model parameters and the corresponding Bayes factors

In this section, we present several priors — i.e., several choices for the density in (1.2) — and derive the expressions of the resulting Bayes factors. In the sequel of the paper, we shall examine the properties (both finite-sample and asymptotic) of the Bayes factors.

3.1. A natural conjugate framework

Both for reasons of computational simplicity and for the interpretability of theoretical results, the most obvious choice for the prior distribution of the parameters is a natural conjugate one. The density in (1.2) is then

given through

$$p(\alpha, \beta_j|\sigma, M_j) = f_N^{k_j+1}((\alpha, \beta_j)|m_{0j}, \sigma^2 V_{0j}), \tag{2.1}$$

which denotes the p.d.f. of a $(k_j + 1)$-variate normal distribution with mean m_{0j} and covariance matrix $\sigma^2 V_{0j}$, and through

$$p(\sigma^{-2}|M_j) = p(\sigma^{-2}) = f_G(\sigma^{-2}|c_0, d_0), \tag{2.2}$$

which corresponds to a Gamma distribution with mean c_0/d_0 and variance c_0/d_0^2 for σ^{-2}. Clearly $m_{0j} \in \Re^{k_j+1}$, V_{0j} a $(k_j + 1) \times (k_j + 1)$ positive-definite symmetric matrix, $c_0 > 0$ and $d_0 > 0$ are prior hyperparameters that still need to be elicited.

This natural conjugate framework greatly facilitates the computation of posterior distributions and Bayes factors. In particular, the marginal likelihood of model M_j computed through (1.8) takes the form

$$l_y(M_j) = f_S^n\left(y|2c_0, X_j m_{0j}, \frac{c_0}{d_0}(I_n - X_j V_{*j} X_j')\right), \tag{2.3}$$

where

$$X_j = (\iota_n : Z_j), \tag{2.4}$$

$$V_{*j} = (X_j' X_j + V_{0j}^{-1})^{-1} \tag{2.5}$$

and $f_S^n(y|v, b, A)$ denotes the p.d.f. of an n-variate Student-t distribution with v degrees of freedom, location vector b (the mean if $v > 1$) and precision matrix A (with covariance matrix $A^{-1}v/(v-2)$ provided $v > 2$) evaluated at y. The Bayes factor for model M_j versus model M_s now takes the form

$$B_{js} = \frac{l_y(M_j)}{l_y(M_s)} = \left(\frac{|V_{*j}||V_{0s}|}{|V_{0j}||V_{*s}|}\right)^{1/2} \times \left\{\frac{2d_0 + (y - X_s m_{0s})'(I_n - X_s V_{*s} X_s')(y - X_s m_{0s})}{2d_0 + (y - X_j m_{0j})'(I_n - X_j V_{*j} X_j')(y - X_j m_{0j})}\right\}^{c_0 + n/2}. \tag{2.6}$$

Generally, the choice of the prior hyperparameters in (2.1)–(2.2) is not a trivial one. The user is plagued by the pitfalls described in Richard (1973), arising if we wish to combine a fixed quantity of subjective prior information on the regression coefficients with little prior information on σ. Richard and Steel (1988, Appendix D) and Bauwens (1991) propose a subjective elicitation procedure for the precision parameter based on the expected fit of the model. See Poirier (1996) for related ideas. In this paper we shall follow the opposite strategy, and instead of trying to elicit more prior information in a situation of incomplete prior

specification, we focus on situations where we have (or wish to use) as little subjective prior knowledge as possible.

3.2. Choosing prior hyperparameters for (α, β_j)

Choosing m_{0j} and V_{0j} can be quite difficult in the absence of prior information. A predictive way of eliciting m_{0j} is through making a prior guess for the n-dimensional response y. Laud and Ibrahim (1996) propose to make such a guess, call it η, taking the information on all the covariates into account and subsequently choose $m_{0j} = (X_j'X_j)^{-1}X_j'\eta$. Our approach is similar in spirit but much simpler: Given that we do not possess a lot of prior information, we consider it very difficult to make a prior guess for n observations taking the covariates for each of these n observations into account. Especially when n is large, this seems like an extremely demanding task. Instead, one could hope to have an idea of the central values of y and make the following prior prediction guess: $\eta = m_1 \iota_n$, which corresponds to

$$m_{0j} = (m_1, 0, \ldots, 0)'. \tag{2.7}$$

Eliciting prior correlations is even more difficult. We adopt the convenient g-prior (Zellner, 1986), which corresponds to taking

$$V_{0j}^{-1} = g_{0j} X_j' X_j \tag{2.8}$$

with $g_{0j} > 0$. From (2.5) it is clear that V_{0j}^{-1} is the prior counterpart of $X_j'X_j$ and, thus, (2.8) implies that the prior precision is a fraction g_{0j} of the precision arising from the sample. This choice is extremely popular, and has been considered, among others by Poirier (1985) and Laud and Ibrahim (1995, 1996). See also Smith and Spiegelhalter (1980) for a closely related idea.

With these hyperparameter choices, the Bayes factor in (2.6) can be written in the following intuitively interpretable way:

$$B_{js} = \left(\frac{g_{0j}}{g_{0j}+1}\right)^{(k_j+1)/2} \left(\frac{g_{0s}+1}{g_{0s}}\right)^{(k_s+1)/2} \times \left(\frac{2d_0 + \frac{1}{g_{0s}+1}\, y' M_{X_s} y + \frac{g_{0s}}{g_{0s}+1}\,(y - m_1\iota_n)'(y - m_1\iota_n)}{2d_0 + \frac{1}{g_{0j}+1}\, y' M_{X_j} y + \frac{g_{0j}}{g_{0j}+1}\,(y - m_1\iota_n)'(y - m_1\iota_n)}\right)^{c_0 + n/2}, \tag{2.9}$$

where

$$y' M_{X_j} y = y'y - y' X_j (X_j'X_j)^{-1} X_j' y \tag{2.10}$$

is the usual sum of squared residuals under model M_j.

Note that the last factor in (2.9) contains a convex combination between the model 'lack of fit' (measured through $y' M_{X_j} y$) and the 'error of our prior

prediction guess' [measured through $(y - m_1\iota_n)'(y - m_1\iota_n)$]. The coefficients of this convex combination are determined by the choice of g_{0j}. The choice of g_{0j} is crucial for obtaining sensible results, as we shall see later. By not choosing g_{0j} through fixing a marginal prior of the regression coefficients, we avoid the natural conjugate pitfall alluded to at the end of Section 3.1. In addition, the g-prior in (2.7)–(2.8) can also lead to a prior that is continuously induced across models, as defined in Poirier (1985), in the sense that the priors for all 2^k models can be derived as the relevant conditionals from the prior of the full model (with $k_j = k$). This will hold as long as g_{0j} does not depend on M_j and we modify the prior in (2.2) so that the shape parameter c_0 becomes model-specific and is replaced by $c_0 + (k - k_j)/2$.

3.3. A non-informative prior for σ

From (2.9) it is clear that the choice of d_0, the precision parameter in the Gamma prior distribution for σ^{-2}, can crucially affect the Bayes factor. In particular, if the value of d_0 is large in relation to the values of $y'M_{X_j}y$ and $(y - m_1\iota_n)'(y - m_1\iota_n)$ the prior will dominate the sample information, which is a rather undesirable property. The impact of d_0 on the Bayes factor also clearly depends on the units of measurement for the data y. In the absence of (or under little) prior information, it is very difficult to choose this hyperparameter value without using the data if we do not want to risk choosing it too large. Even using prior ideas about fit does not help; Poirier (1996) shows that the population analog of the coefficient of determination (R^2) does not have any prior dependence on c_0 or d_0. Use of the information in the response variable was proposed, e.g., by Raftery (1996) and Raftery et al. (1997) but, as we already mentioned, we prefer to avoid this situation. Instead we propose the following:

Since the scale parameter σ appears in all the models entertained, we can use the improper prior distribution with density

$$p(\sigma) \propto \sigma^{-1}, \tag{2.11}$$

which is the widely accepted non-informative prior distribution for scale parameters. Note that we have assumed a common prior distribution for σ across models. This practice is often followed in the literature — see e.g., Mitchell and Beauchamp (1988) and Raftery et al. (1997) — and leads to procedures with good operating characteristics. It is easy to check that the improper prior in (2.11) results in a proper posterior (and thus allows for a Bayesian analysis) as long as $y \neq m_1\iota_n$.

The distribution in (2.11) is the only one that is invariant under scale transformations (induced by, e.g., a change in the units of measurement) and is the limiting distribution of the Gamma conjugate prior in (2.2) when both

d_0 and c_0 tend to zero. This leads to the Bayes factor

$$B_{js} = \left(\frac{g_{0j}}{g_{0j}+1}\right)^{(k_j+1)/2}\left(\frac{g_{0s}+1}{g_{0s}}\right)^{(k_s+1)/2} \times\left(\frac{\frac{1}{g_{0s}+1}\, y'M_{X_s}y + \frac{g_{0s}}{g_{0s}+1}\,(y-m_1\iota_n)'(y-m_1\iota_n)}{\frac{1}{g_{0j}+1}\, y'M_{X_j}y + \frac{g_{0j}}{g_{0j}+1}\,(y-m_1\iota_n)'(y-m_1\iota_n)}\right)^{n/2}, \tag{2.12}$$

where we have avoided the influence of the hyperparameter values c_0 and d_0.

3.4. A non-informative prior for the intercept

In (2.12) there are two subjective elements that still remain, namely the choices of g_{0j} and of m_1, where $m_1\iota_n$ is our prior guess for y. It is clear from (2.12) that the choice of m_1 can have a non-negligible impact on the actual Bayes factor and, under absence of prior information, it is extremely difficult to successfully elicit m_1 without using the data. The idea that we propose here is in line with our solution for the prior on σ: since all the models have an intercept, take the usual non-informative improper prior for a location parameter with constant density. This avoids the difficult issue of choosing a value for m_1.

This setup takes us outside the natural conjugate framework, since our prior for (α, β_j) no longer corresponds to (2.1). Without loss of generality, we assume that

$$\iota_n' Z = 0, \tag{2.13}$$

so that the intercept is orthogonal to all the regressors. This is immediately achieved by subtracting the corresponding mean from each of them. Such a transformation only affects the interpretation of the intercept α, which is typically not of primary interest. In addition, the prior that we next propose for α is not affected by this transformation. We now consider the following prior density for (α, β_j):

$$p(\alpha) \propto 1, \tag{2.14}$$

$$p(\beta_j|\sigma, M_j) = f_N^{k_j}(\beta_j|0, \sigma^2(g_{0j}Z_j'Z_j)^{-1}). \tag{2.15}$$

Through (2.14)–(2.15) we assume the same prior distribution for α in all of the models and a g-prior distribution for β_j under model M_j. We again use the non-informative prior described in (2.11) for σ. Existence of a proper posterior distribution is now achieved as long as the sample contains at least two different observations. The Bayes factor for M_j versus M_s now is

$$B_{js} = \left(\frac{g_{0j}}{g_{0j}+1}\right)^{k_j/2}\left(\frac{g_{0s}+1}{g_{0s}}\right)^{k_s/2} \left(\frac{\frac{1}{g_{0s}+1}\, y'M_{X_s}y + \frac{g_{0s}}{g_{0s}+1}\,(y-\bar{y}\iota_n)'(y-\bar{y}\iota_n)}{\frac{1}{g_{0j}+1}\, y'M_{X_j}y + \frac{g_{0j}}{g_{0j}+1}\,(y-\bar{y}\iota_n)'(y-\bar{y}\iota_n)}\right)^{(n-1)/2} \tag{2.16}$$

if $k_j \geqslant 1$ and $k_s \geqslant 1$, where $\bar{y} = \iota_n' y/n$. If one of the latter two quantities, e.g., k_j, is zero (which corresponds to the model with just the intercept), the Bayes factor is simply obtained as the limit of B_{js} in (2.16) letting g_{0j} tend to infinity.

Note the similarity between the expression in (2.16) and (2.12), where we had adopted a (limiting) natural conjugate framework. When we are non-informative on the intercept — see (2.16) — we lose, as it were, one observation (n becomes $n-1$) and one regressor ($k_j + 1$ becomes k_j). But the most important difference is that our subjective prior guess m_1 is now replaced by $\bar{y}$, which seems quite reasonable and avoids the sensitivity problems alluded to before. Thus, we shall, henceforth, focus on the prior given by the product of (2.11), (2.14) and (2.15), leading to the Bayes factor in (2.16). Note that only the scalar g_{0j} remains to be chosen. This choice will be inspired by properties of the posterior model probabilities and predictive ability.

4. The simulation experiment

4.1. Introduction

In this section we describe a simulation experiment to assess the performance of different choices of g_{0j} in finite sampling. Among other things, we will compute posterior model probabilities and evaluate predictive ability under several choices of g_{0j}. Our results will be derived under a Uniform prior on the model space $\mathcal{M}$. Thus, the Bayesian model will be given through (1.1), together with the prior densities in (2.11), (2.14) and (2.15), and

$$P(M_j) = p_j = 2^{-k}, \quad j = 1, \ldots, 2^k. \tag{3.1}$$

Adopting (3.1) (as in the examples of George and McCulloch, 1993, Smith and Kohn, 1996; Raftery et al., 1997) is another expression of lack of substantive prior information, but we stress that there might be cases in which other choices are more appropriate. One possibility would be to downweigh models with many regressors. Chipman (1996) examines prior structures that can be used to accommodate general relations between regressors.

In all, we will analyse three models that are chosen to reflect a wide variety of situations. Creating the design matrix of the simulation experiment for the first two models follows Example 5.2.2 in Raftery et al. (1997). We generate an $n \times k$ ($k = 15$) matrix R of regressors in the following way: the first ten columns in R, denoted by $(r_{(1)}, \ldots, r_{(10)})$ are drawn from independent standard Normal distributions, and the next five columns $(r_{(11)}, \ldots, r_{(15)})$ are constructed from

$$(r_{(11)}, \ldots, r_{(15)}) = (r_{(1)}, \ldots, r_{(5)})(0.3\ 0.5\ 0.7\ 0.9\ 1.1)'(1\ 1\ 1\ 1\ 1) + E, \tag{3.2}$$

where E is an $n \times 5$ matrix of independent standard normal deviates. Note that (3.2) induces a correlation between the first five regressors and the last five

regressors. The latter takes the form of small to moderate correlations between $r_{(i)}$, $i = 1, \ldots, 5$, and $r_{(11)}, \ldots, r_{(15)}$ (the theoretical correlation coefficients increase from 0.153 to 0.561 with i) and somewhat larger correlations between the last five regressors (theoretical values 0.740).[3] After generating R, we demean each of the regressors, thus leading to a matrix $Z = (z_{(1)}, \ldots, z_{(15)})$ that fulfills (2.13). A vector of n observations is then generated according to one of the models

$$\text{Model 1:} \quad y = 4\iota_n + 2z_{(1)} - z_{(5)} + 1.5z_{(7)} + z_{(11)} + 0.5z_{(13)} + \sigma\varepsilon, \tag{3.3}$$

$$\text{Model 2:} \quad y = \iota_n + \sigma\varepsilon, \tag{3.4}$$

where the n elements of ε are i.i.d. standard normal and $\sigma = 2.5$. In our simulations, n takes the values 50, 100, 500, 1000, 10,000 and 100,000. Whereas Model 1 is meant to capture a more or less realistic situation where one third of the regressors intervene (the theoretical 'R^2' is 0.55 for this model), Model 2 is an extreme case without any relationship between predictors and response. A 'null model' similar to the latter was analysed in Freedman (1983) using a classical approach and in Raftery et al. (1997) through Bayesian model averaging.

The third model considers widely varying values for k, namely $k = 4$, 10, 20 and 40. For each choice of k, a similar setup to Example 4.2 in George and McCulloch (1993) was followed. In particular, we generate k regressors as $r_{(i)} = r^*_{(i)} + e$, $i = 1, \ldots, k$ where each $r^*_{(i)}$ and e are n-dimensional vectors of independent standard normal deviates. This induces a pairwise theoretical correlation of 0.5 between all regressors. Again, Z will denote the $n \times k$ matrix of demeaned regressors. The n observations are then generated through

$$\text{Model 3:} \quad y = \iota_n + \sum_{h=1}^{k/2} z_{(k/2+h)} + \sigma\varepsilon, \tag{3.5}$$

where the n elements of ε are again i.i.d. standard normal and now $\sigma = 2$. Choices for n will be restricted to 100 and 1000, values of particular practical interests for many applications. The theoretical 'R^2' varies from 0.43 (for $k = 4$) to 0.98 (for $k = 40$) in this model, covering a reasonable range of values.

4.2. Choices for g_{0j}

We consider the following nine choices:

Prior a: $g_{0j} = 1/n$. This prior roughly corresponds to assigning the same amount of information to the conditional prior of β as is contained in one observation. Thus, it is in the spirit of the 'unit information priors' of Kass and

[3] This correlation structure differs from the one reported in Raftery et al. (1997), which seems in conflict with (3.2).

Wasserman (1995) and the g-prior (using a Cauchy prior on β given σ) used in Zellner and Siow (1980). Kass and Wasserman (1995) state that the intrinsic Bayes factors of Berger and Pericchi (1996) and the fractional Bayes factors of O'Hagan (1995) can in some cases yield similar results to those obtained under unit information priors.

Prior b: $g_{0j} = k_j/n$. Here we assign more information to the prior as we have more regressors in the model, i.e., we induce more shrinkage in β_j (to the prior mean of zero) as the number of regressors grows.

Prior c: $g_{0j} = k^{1/k_j}/n$. Now prior information decreases with the number of regressors in the model.

Prior d: $g_{0j} = \sqrt{1/n}$. This is an intermediate case, where we choose a smaller asymptotic penalty term for large models than in the Schwarz criterion (see (A.19) in the appendix), which corresponds to priors a–c.

Prior e: $g_{0j} = \sqrt{k_j/n}$. As in prior b, we induce more shrinkage as the number of regressors grows.

Prior f: $g_{0j} = 1/(\ln n)^3$. Here we choose g_{0j} so as to mimic the Hannan–Quinn criterion in (A.20) with $C_{HQ} = 3$ as n becomes large.

Prior g: $g_{0j} = \ln(k_j + 1)/\ln n$. Now g_{0j} decreases even slower with sample size and we have asymptotic convergence of $\ln B_{js}$ to the Hannan–Quinn criterion with $C_{HQ} = 1$.

Prior h: $g_{0j} = \delta\gamma^{1/k_j}/(1 - \delta\gamma^{1/k_j})$. This choice was suggested by Laud and Ibrahim (1996), who use a natural conjugate prior structure, subjectively elicited through predictive implications. In applications, they propose to choose $\gamma < 1$ (so that g_{0j} increases with k_j) and δ such that $g_{0j}/(1 + g_{0j}) \in [0.10, 0.15]$ (the weight of the 'prior prediction error' in our Bayes factors); for k_j ranging from 1 to 15, we cover this interval with the values $\gamma = 0.65$, $\delta = 0.15$.

Prior i: $g_{0j} = 1/k^2$. This prior is suggested by the risk inflation criterion (RIC) of Foster and George (1994) (see comment below).

From the results in the appendix, priors a–g all lead to consistency, in the sense of asymptotically selecting the correct model, whereas priors h–i do not in general. In addition, the log Bayes factors obtained under priors a–c behave asymptotically like the Schwarz criterion, whereas those obtained under priors f and g behave like the Hannan–Quinn criterion, with $C_{HQ} = 3$ and 1, respectively. Priors d and e provide an intermediate case in terms of asymptotic penalty for large models.

George and Foster (1997) show that in a linear regression model with a g-prior on the regression coefficients and known σ^2 the selection of the model with highest posterior probability is equivalent (for any sample size) to choosing the model with the highest value for the RIC provided we take $g_{0j} = 1/k^2$. Whereas our model is different (no g-prior on the intercept and unknown σ^2), we still think it is interesting to examine this choice for g_{0j} in our context and adopt it as prior i. In the same context, George and Foster (1997) show that AIC corresponds to choosing $g_{0j} = 0.255$ and BIC (Schwarz) to $g_{0j} = 1/n$. Thus, we

can roughly compare AIC to prior h where g_{0j} takes the largest values, and the relationship between the Schwarz criterion and prior a goes beyond mere asymptotics.

4.3. Predictive criteria

Clearly, if we generate the data from some known model, we are interested in recovering that model with the highest possible posterior probability for each given sample size n. However, in practical situations with real data, we might be more interested in predicting the observable, rather than uncovering some 'true' underlying structure. This is more in line with the Bayesian way of thinking, where models are mere 'windows' through which to view the world (Poirier, 1988), but have no inherent meaning in terms of characteristics of the real world. See also Dawid (1984) and Geisser and Eddy (1979).

Forecasting is conducted conditionally upon the regressors, so we will generate q k-dimensional vectors $z_f, f = 1, \ldots, q$, given which we will predict the observable y. In empirical applications, z_f will typically be constructed from some original value r_f from which we subtract the mean of the raw regressors R in the sample on which inference is based. This ensures that the interpretation of the regression coefficients in posterior and predictive inference is compatible.

In this subsection, it will prove useful to make the conditioning on the regressors in z_f and Z explicit in the notation. The out-of-sample predictive distribution for $f = 1, \ldots, q$ will be characterized by

$$p(y_f|z_f, y, Z) = \sum_{j=1}^{2^k} f_S^1\Bigg(y_f|n-1,\ \bar{y} + \frac{1}{g_{0j}+1}\, z'_{f,j}\beta_j^*, \times \frac{n-1}{d_j^*}\left\{1 + \frac{1}{n} + \frac{1}{g_{0j}+1}\, z'_{f,j}(Z_j'Z_j)^{-1}z_{f,j}\right\}^{-1}\Bigg) P(M_j|y, Z), \tag{3.6}$$

where $\bar{y}$ is based on the inference sample $y = (y_1, \ldots, y_n)'$, $z_{f,j}$ groups the j elements of z_f corresponding to the regressors in M_j, $\beta_j^* = (Z_j'Z_j)^{-1}Z_j'y$ and

$$d_j^* = \frac{1}{g_{0j}+1}\, y'M_{X_j}y + \frac{g_{0j}}{g_{0j}+1}\, (y - \bar{y}\iota_n)'(y - \bar{y}\iota_n). \tag{3.7}$$

The term in (3.6) corresponding to the model with only the intercept is obtained by letting the corresponding g_{0j} tend to infinity.

The log predictive score is a proper scoring rule introduced by Good (1952). Some of its properties are discussed in Dawid (1986). For each value of z_f we shall generate a number, say v, of responses from the underlying true model ((3.3), (3.4) or (3.5)) and base our predictive measure on (3.6) evaluated in these

out-of-sample observations $y_{f_1}, \ldots, y_{f_v}$, namely

$$LPS(z_f, y, Z) = -\frac{1}{v}\sum_{i=1}^{v} \ln p(y_{f_i}|z_f, y, Z). \tag{3.8}$$

It is clear that a smaller value of $LPS(z_f, y, Z)$ makes a Bayes model (thus, in our context, a prior choice for g_{0j}) preferable. Madigan, et al. (1995) give an interpretation for differences in log predictive scores in terms of one toss with a biased coin.

More formally, the criterion in (3.8) can be interpreted as an approximation to the expected loss with a logarithmic rule, which is linked to the well-known Kullback–Leibler criterion. The Kullback–Leibler divergence between the actual sampling density $p(y_f|z_f)$ in (3.3), (3.4) or (3.5) and the out-of-sample predictive density in (3.6) can be written as

$$KL\{p(y_f|z_f), p(y_f|z_f, y, Z)\} = \int_{\Re} \{\ln p(y_f|z_f)\} p(y_f|z_f)\, \mathrm{d}y_f - \int_{\Re} \{\ln p(y_f|z_f, y, Z)\} p(y_f|z_f)\, \mathrm{d}y_f, \tag{3.9}$$

where the first integral is the negative entropy of the sampling density, and the second integral can be seen as a theoretical counterpart of (3.8) for a given value of z_f. This latter integral can easily be shown to be finite in our particular context and is now approximated by averaging over v values for y_{f_i} given a particular vector of regressors z_f. For the normal sampling model used here, the negative entropy is given by $-\frac{1}{2}\{\ln(2\pi\sigma^2) + 1\} = -2.335$ for our choice of σ in (3.3) and (3.4), and -2.112 for (3.5), regardless of z_f. By the non-negativity of the Kullback–Leibler divergence, this constitutes a lower bound for $LPS(z_f, y, Z)$ of 2.335 or 2.112.

We can also investigate the calibration of the predictive and compare the entire predictive density function in (3.6) with the known sampling distribution of the response in (3.3), (3.4) or (3.5) given a particular (fixed) set of regressor variables. The fact that such predictions are, by the very nature of our regression model, conditional upon the regressors does complicate matters slightly. We cannot simply compare the sampling density averaged over different values of z_f with the averaged predictive density function. It is clearly crucial to identify predictives with the value of z_f they condition on. Predicting correctly 'on average' can mask arbitrarily large errors in conditional predictions, as long as they compensate each other. For Model 1, we shall graphically present comparisons of the sampling density and the predictive density for three key values of z_f within our sample of q predictors: the one leading to the smallest mean of the sampling model in (3.3), the one leading to the median value and the one giving

rise to the largest value. In addition, we have computed quantiles of *LPS* and of predictive coverage over the different values of z_f as well. These latter measures of predictive performance naturally compare each predictive with the corresponding sampling distribution (i.e., taking the value of z_f into account), so that an overall measure can readily be computed.

5. Simulation results

5.1. Convergence and implementation

The implementation of the simulation study described in the previous section will be conducted through the MC^3 methodology mentioned in Section 1. This Metropolis algorithm generates a new candidate model, say M_j, from a Uniform distribution over the subset of $\mathscr{M}$ consisting of the current state of the chain, say M_s, and all models containing either one regressor more or one regressors less than M_s. The chain moves to M_j with probability $\min(1, B_{js})$, where B_{js} is the Bayes factor in (2.16).

In order to evaluate the posterior model probabilities we can simply count the relative frequencies of model visits in the induced Markov chain. A somewhat more interesting alternative to this strategy is to use the actual Bayes factors, already computed in running the chain to compare all visited models. Since the number of visited models is typically a small subset of the total number of possible models, this method is feasible. This idea is called 'window estimation' in Clyde et al. (1996) and Lee (1996) mentions it as 'Bayesian random search' (BARS). The generated chain is then effectively only used to indicate which models should be considered in computing Bayes factors. All other (non-visited) models will implicitly be assumed to have zero posterior probability. This has two advantages: firstly, it is clearly more precise than relative frequencies, since the Bayes factors in (2.16) are exact and do not require any ergodic properties. Clyde et al. (1996) provide some empirical support for this claim. Secondly, comparing empirical relative frequencies with exact Bayes factors will give a good indication of the convergence of the chain. We shall report results based on Bayes factors, but we ran the chain for long enough to get virtually the same answers with empirical model frequencies. This was obtained with 50,000 recorded draws after a burn-in of 20,000 draws. A useful diagnostic to assess convergence of the Markov chain is the correlation coefficient of the model probabilities based on the exact Bayes factors computed through (2.16) and the relative frequencies of model visits. In our simulation experiment, this correlation coefficient was typically above 0.99. If we are interested in estimating how much of the total probability mass we have captured in the visited models, we can compare exact Bayes factors and relative frequencies of a prespecified subset of models in the way indicated in George and McCulloch (1997, Section 4.5).

In order to avoid results depending on the particular sample analysed, we have generated 100 independent samples (y, Z) according to the setup described in Section 4. Frequently, results will be presented in the form of either means and standard deviations or quantiles computed over these 100 samples. Sample sizes (i.e., values of n) used in the simulation are as indicated in Section 4.1. Furthermore, we generate $q = 19$ different vectors of regressors z_f for the forecasts of Models 1 and 3, whereas $q = 5$ for Model 2. For each of these values of the vector z_f, $v = 100$ out-of-sample observations will be generated.[4]

Due to space limitations, we will only present the most relevant findings in detail, and will briefly summarize the remaining results.

5.2. Posterior model inference

5.2.1. Results under Model 1

One of the indicators of the performance of the Bayesian methodology is the posterior probability assigned to the model that has generated the data. Ideally, one would want this probability to be very high for small or moderate values of n that are likely to occur in practice. Table 1 presents the means and standard deviations across the 100 samples of (y, Z) for the posterior probability of the true model (Model 1). Columns correspond to the six sample sizes used and rows order the different priors introduced in Section 4.2. In order to put these results in a better perspective, note that the prior model probability of each of the 2^{15} possible models is equal and amounts to 3.052×10^{-5}. We know from the theoretical results in the appendix (Section A.1) that priors a–g are consistent. From Section A.2, we remain inconclusive about consistency under prior h, and we know prior i will asymptotically allocate mass to models that nest the true model. Our simulation results suggest that consistency holds for prior h (but, indeed, not for prior i) in our particular example. It is clear from Table 1 that the posterior probability of Model 1 varies greatly in finite samples. Whereas prior e already performs very well for $n = 1000$, getting average probabilities of the correct model upwards of 0.97, prior d only obtains a probability of 0.36 with a sample as large as 100,000. This result is all the more striking, since the asymptotic behaviour with both priors is the same, and they are clearly very related. This underlines the inherent sensitivity of Bayes factors to the particular choice of g_{0j}. In view of this poor performance in a critical issue, we will often

[4] As such a simulation study is quite CPU demanding, we put a good deal of emphasis on efficient coding and speed of execution. We coded in standard Fortran 77, and we used stacks to store information pertaining to evaluated models in order to reduce the number of calculations. On a PowerMacintosh 7600, each 20,000–50,000 chain for Model 1 would take an average (over priors) time in seconds of: 209, 58, 15, 5, 18, and 117; for $n = 50$, 100, 500, 1000, 10,000 and 100,000. Since the number of visited models (and thus, the number of marginal likelihood calculations) will typically decrease with n, CPU requirements are not monotone in sample size.

Table 1
Model 1: Means and Stds of the posterior probability of the true model

n	50		100		500		1000		10,000		100,000	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	0.0128	0.0197	0.0575	0.0618	0.4013	0.1521	0.5293	0.1401	0.8111	0.0928	0.9254	0.0760
b	0.0066	0.0091	0.0332	0.0338	0.3083	0.1337	0.4407	0.1373	0.7601	0.1064	0.9048	0.0841
c	0.0110	0.0159	0.0519	0.0533	0.3603	0.1441	0.4860	0.1374	0.7853	0.0999	0.9145	0.0804
d	0.0028	0.0029	0.0093	0.0082	0.0541	0.0351	0.0786	0.0425	0.2051	0.0874	0.3625	0.1204
e	0.0029	0.0026	0.0205	0.0188	0.7487	0.1745	0.9730	0.0196	1.0000	0.0000	1.0000	0.0000
f	0.0141	0.0223	0.0586	0.0616	0.2948	0.1307	0.3610	0.1251	0.5139	0.1327	0.5981	0.1327
g	0.0020	0.0014	0.0128	0.0107	0.4805	0.3793	0.7762	0.3421	1.0000	0.0000	1.0000	0.0000
h	0.0026	0.0026	0.0069	0.0056	0.0728	0.0376	0.2773	0.0864	1.0000	0.0000	1.0000	0.0000
i	0.0295	0.0446	0.0961	0.1014	0.2857	0.1284	0.3061	0.1103	0.3094	0.1115	0.3067	0.1161

not give explicit results for prior d in the sequel. Prior i does very well for small sample sizes, but then seems to taper off for values of $n \geqslant 1000$ at a moderate probability for the true model around 0.3. Apart from the absolute probability of the correct model, it is also important to examine how much posterior weight is assigned to Model 1 relative to other models. Therefore, Table 2 presents quartiles of the ratio between the posterior probability of the correct model and the highest posterior probability of any other model. It is clear that in most cases this ratio tends to be far above unity, which is reassuring as it tells us that the most favoured model will still be the correct one, even though it may not have a lot of posterior mass attached to it. For example, with $n = 50$ prior g only leads to a mean posterior probability of Model 1 of 0.002 but still favours the correct model to the next best. In fact, the correct model is always favoured in at least 75 of the 100 samples, even for small sample sizes. Note that this compares favourably to results in George and McCulloch (1993).

Table 3 records means and standard deviations of the number of visited models in the 50,000 recorded draws of the chain in model space (i.e., after the burn-in). Given that the model that generated the data is one of the $2^{15} = 32{,}768$ possible models examined, we would want this to be as small as possible. For $n = 50$ it is clear that the sample information is rather weak, allowing the chain to wander around and visit many models: as much as around a quarter of the total amount of models for prior g, and never less than 3.5% on average (prior i). The sampler visits fewer models as n increases, and for $n = 1000$ we already have very few visited models for prior g in particular and also for e. Of course, depending on the field of application, 1000 observations may well be considered quite a large sample. When 10,000 observations are available, that is enough to make the sampler stick to one model (the correct one) for priors e, g and h. Surprisingly, whereas prior h still leads to very erratic behaviour of the sampler with $n = 1000$, it never fails to put all the mass on the correct model for the larger sample sizes. Even with 100,000 observations, priors d and f (though consistent) still make the sampler visit 240 and 110 models on average. For prior i this is as high as 287 models.

Table 4 indicates in what sense the different Bayesian models tend to err if they assign posterior probability to alternative sampling models. In particular, Table 4 presents the means and standard deviations of the posterior probabilities of including each of the regressors. As we know from (3.3), Model 1 contains regressors 1, 5, 7, 11 and 13 (indicated with arrows in Table 4). To save space, we shall only report these results for $n = 50$ and $n = 1000$, and we will not include prior c (for which results were virtually identical to prior a) and prior d. When $n = 50$, regressors $z_{(1)}$ and $z_{(7)}$ are almost always included. Since they are (almost) orthogonal to the other regressors, and their regression coefficients are rather large in absolute value, this is not surprising. Regressor $z_{(11)}$ is only correlated with $z_{(13)}$ and is still often included. The most difficult are regressors 5 and 13, which are positively correlated, and have relatively small regression

Table 2
Model 1: Quartiles of ratio of posterior probabilities; true model versus best among the rest

n	50			100			500			1000			10,000		
Prior	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
a	1.5	3.2	6.3	3.0	5.8	8.6	3.5	13.5	20.0	9.1	19.0	29.8	27.6	67.8	90.5
b	1.1	2.6	4.2	2.2	4.1	6.2	5.5	10.8	15.3	9.6	16.6	22.1	16.6	36.3	66.3
c	1.2	3.4	5.8	2.0	5.1	8.3	7.6	13.7	18.5	7.0	13.8	22.9	26.3	55.9	73.6
e	1.6	2.7	3.5	1.9	3.8	5.8	18.8	53.6	73.6	226.5	416.4	629.4	∞	∞	∞
f	1.7	4.0	7.0	2.0	4.4	8.7	5.1	10.9	14.1	4.7	9.2	16.5	9.7	19.7	24.9
g	1.4	2.3	3.3	1.4	3.9	5.1	6.3	53.9	250.2	11.4	238.7	3625.8	∞	∞	∞
h	1.2	2.3	2.8	1.9	2.8	3.9	2.9	4.9	5.7	5.9	10.6	12.9	∞	∞	∞
i	1.1	4.3	9.8	2.1	6.5	12.5	4.3	8.8	12.8	4.8	8.4	13.0	6.0	10.9	14.6

Table 3
Model 1: Means and Stds of number of models visited

n Prior	50		100		1000		10,000		100,000	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	2230	637	1123	290	134	29	49	11	20	5
b	4478	1347	1994	615	206	46	66	16	25	7
c	2475	711	1252	317	148	32	53	11	21	5
e	7159	1549	2810	838	15	4	1	0	1	0
f	2056	596	1158	301	237	47	151	37	110	24
g	8677	1608	3555	1204	3	1	1	0	1	0
h	5480	1353	3322	809	654	89	1	0	1	0
i	1176	451	758	222	288	51	288	55	287	53

coefficients of opposite signs. The posterior probabilities of including regressors not contained in the correct model are all relatively small. Note that this is exactly where prior i excels, as the posterior probabilities of including incorrect regressors are much smaller than for the other priors. What is not clearly exemplified by Table 4 is that most priors tend to choose alternatives that are nested by Model 1 for small sample sizes, with the exception of priors d and h, which put considerable posterior mass on models that nest the correct sampling model. Table 4 informs us that for $n = 1000$ the correct regressors are virtually always included. Only prior g has a tendency to choose models that are nested by Model 1. For the other priors there remain small probabilities of incorrectly including extra regressors (the smallest for prior e and the largest for priors d and h). Alternative models tend to nest the correct model for all priors, except prior g, with this and larger sample sizes.

5.2.2. Results under Model 2

Let us now briefly present the results when the data are generated according to Model 2 in (3.4), the null model. Table 5 presents means and standard deviations of the posterior probability of the null model. It is clear that this is not an easy task (see also the discussion in Freedman, 1983; Raftery et al., 1997) and most priors lead to small probabilities of selecting the correct model. Overall, prior i does best for small sample sizes (followed by priors a and f), whereas larger sample sizes are most favourable to priors a and b. The behaviour of prior i is quite striking: it appears that sample size has very little influence of the posterior probability of the correct model, making it a clear winner for values of $n \leqslant 100$. Note that the other prior where g_{0j} does not depend on sample size, prior h, also has a posterior probability of the null model that is roughly constant in n (but now at a very low level). Despite the rather small posterior probabilities of the null model, the latter is still typically favoured over the second best model. This is evidenced by Table 6, where the three quartiles of

Table 4
Model 1: Means and Stds of posterior probabilities of including each regressor

Prior Reg.	a		b		e		f		g		h		i	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$n = 50$														
→ 1	0.98	0.07	0.98	0.07	0.97	0.09	0.98	0.07	0.96	0.09	0.98	0.06	0.98	0.10
2	0.22	0.17	0.29	0.14	0.33	0.10	0.21	0.17	0.35	0.09	0.36	0.13	0.13	0.13
3	0.25	0.18	0.31	0.14	0.35	0.11	0.24	0.18	0.37	0.10	0.38	0.13	0.15	0.12
4	0.27	0.19	0.33	0.15	0.37	0.11	0.25	0.19	0.39	0.10	0.40	0.13	0.21	0.21
→ 5	0.42	0.27	0.44	0.22	0.43	0.15	0.40	0.27	0.43	0.13	0.50	0.19	0.41	0.30
6	0.22	0.16	0.28	0.13	0.32	0.09	0.21	0.15	0.34	0.08	0.35	0.13	0.10	0.10
→ 7	0.94	0.14	0.94	0.13	0.90	0.15	0.94	0.15	0.87	0.15	0.94	0.12	0.86	0.22
8	0.22	0.16	0.29	0.14	0.32	0.10	0.21	0.15	0.34	0.08	0.36	0.13	0.13	0.14
9	0.21	0.14	0.28	0.12	0.32	0.08	0.20	0.14	0.34	0.07	0.35	0.11	0.12	0.11
10	0.21	0.14	0.28	0.12	0.32	0.08	0.20	0.13	0.34	0.07	0.35	0.11	0.11	0.08
→ 11	0.82	0.25	0.81	0.22	0.76	0.19	0.81	0.25	0.74	0.18	0.82	0.20	0.73	0.30
12	0.24	0.19	0.30	0.15	0.34	0.11	0.23	0.19	0.36	0.09	0.37	0.14	0.15	0.15
→ 13	0.39	0.27	0.43	0.23	0.44	0.18	0.38	0.27	0.45	0.15	0.49	0.20	0.33	0.28
14	0.27	0.22	0.32	0.19	0.36	0.13	0.25	0.22	0.37	0.11	0.39	0.16	0.15	0.16
15	0.22	0.15	0.28	0.12	0.33	0.08	0.21	0.15	0.35	0.07	0.36	0.11	0.15	0.16
$n = 1000$														
→ 1	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
2	0.07	0.11	0.09	0.12	0.00	0.01	0.11	0.13	0.00	0.00	0.16	0.10	0.11	0.07
3	0.06	0.08	0.08	0.08	0.00	0.01	0.09	0.09	0.00	0.00	0.15	0.07	0.11	0.08
4	0.05	0.06	0.07	0.07	0.00	0.01	0.08	0.08	0.00	0.00	0.14	0.06	0.10	0.07
→ 5	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.88	0.26	1.00	0.00	1.00	0.00
6	0.07	0.08	0.09	0.09	0.00	0.00	0.10	0.10	0.00	0.00	0.16	0.08	0.11	0.11
→ 7	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
8	0.06	0.07	0.08	0.08	0.00	0.00	0.10	0.10	0.00	0.00	0.15	0.08	0.10	0.06
9	0.06	0.06	0.08	0.08	0.00	0.00	0.10	0.09	0.00	0.00	0.15	0.07	0.11	0.09
10	0.06	0.07	0.08	0.08	0.00	0.00	0.10	0.09	0.00	0.00	0.15	0.07	0.12	0.13
→ 11	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
12	0.06	0.10	0.08	0.10	0.00	0.01	0.09	0.10	0.00	0.00	0.15	0.09	0.11	0.13
→ 13	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.78	0.34	1.00	0.00	1.00	0.00
14	0.05	0.04	0.07	0.05	0.00	0.00	0.08	0.06	0.00	0.00	0.14	0.05	0.13	0.14
15	0.06	0.07	0.07	0.08	0.00	0.00	0.09	0.09	0.00	0.00	0.14	0.07	0.11	0.07

Table 5
Model 2: Means and Stds of the posterior probability of the true model

n	50		100		500		1000		10,000		100,000	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	0.0320	0.0269	0.0722	0.0494	0.2707	0.1338	0.3812	0.1543	0.7199	0.1346	0.8995	0.0529
b	0.0021	0.0028	0.0114	0.0124	0.1394	0.1055	0.2606	0.1506	0.6910	0.1441	0.8960	0.0568
c	0.0099	0.0081	0.0238	0.0157	0.0994	0.0505	0.1494	0.0715	0.4148	0.1201	0.7080	0.1009
e	0.0003	0.0003	0.0006	0.0006	0.0034	0.0036	0.0066	0.0066	0.0427	0.0300	0.1570	0.0938
f	0.0407	0.0322	0.0764	0.0492	0.1787	0.0959	0.2216	0.1050	0.3569	0.1220	0.4733	0.1235
g	0.0001	0.0002	0.0002	0.0002	0.0005	0.0005	0.0005	0.0006	0.0009	0.0008	0.0014	0.0015
h	0.0010	0.0012	0.0011	0.0012	0.0016	0.0015	0.0014	0.0014	0.0013	0.0011	0.0014	0.0014
i	0.1599	0.0862	0.1658	0.0842	0.1833	0.0870	0.1790	0.0891	0.1780	0.0886	0.1862	0.0815

Table 6
Model 2: Quartiles of ratio of posterior probabilities; true model versus best among the rest

n	50			100			1000			10,000			100,000		
Prior	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
a	2.4	4.4	7.0	3.1	6.2	9.2	6.0	16.1	24.8	30.0	60.0	86.2	78.7	202.4	272.1
b	2.1	5.2	7.0	3.7	6.7	9.7	7.4	22.1	28.9	18.4	45.7	79.6	64.1	153.8	253.2
c	1.3	1.8	2.1	1.2	2.2	2.7	2.0	4.1	7.1	6.3	15.2	21.8	16.1	40.3	70.2
e	1.1	2.1	2.8	1.2	2.5	3.2	2.2	3.9	5.2	3.1	6.6	9.4	6.1	11.2	16.7
f	2.4	5.3	7.5	3.5	6.8	9.3	4.9	11.4	17.5	8.1	15.7	25.3	12.9	26.9	36.2
g	0.5	1.5	2.6	0.6	1.6	2.6	0.9	2.3	3.2	1.3	2.6	3.5	1.8	3.4	4.2
h	1.2	2.3	3.2	1.3	2.5	3.2	1.3	2.7	3.2	1.7	2.7	3.5	1.6	2.5	3.1
i	4.2	8.2	13.4	5.5	10.7	13.7	4.6	9.8	14.3	6.3	10.6	14.1	4.3	9.2	14.4

the ratio of the posterior probabilities of Model 2 and the best other model are presented. Only prior g for $n \leqslant 1000$ leads to a first quartile below unity. Clearly, prior i does best for $n \leqslant 100$ and prior a does best for larger n. The difficulty of pinning down the correct (null) model can also be inferred from the number of visited models (not presented in detail). Some priors (like d, e, g and h) make the chain wander a lot for small sample sizes. Priors g and h retain this problematic behaviour even for sample sizes as large as 100,000. Interestingly, whereas prior g leads to (very slow) improvements as n increases, the bad behaviour with prior h seems entirely unaffected by sample size, as remarked above. Of course, we know from the theory in the appendix that priors h and i do not lead to consistent Bayes factors in this case. Whereas prior h visits on average about 12,500 models for any sample size, prior i visits around one tenth of that. The number of models visited is relatively small for priors i and a, which seem to emerge as the clear winners from the posterior results under Model 2, respectively, for small and large values of n.

5.2.3. Results under Model 3

Now the setup of the experiment is slightly different, as we contrast various model sizes (in terms of k) and only two sample sizes. We would expect that the task of identifying the true model becomes harder as k increases, since the total number of models in $\mathscr{M}$ increases dramatically from 16 (for $k = 4$) to 1.1×10^{12} (for $k = 40$). Of course, this may be partly offset by the fact that σ remains the same, so that the theoretical coefficient of determination grows with k. Tables 7 and 8 present the posterior probability of Model 3 in (3.5) with $n = 100$ and 1000 observations, respectively. The first thing to notice from Table 7 is that priors where g_{0j} increases with k_j (priors b, e, g and h) suffer a large drop in performance as the true model (and k) become large ($k \geqslant 20$ and even $k = 10$ for

Table 7
Model 3: Means and Stds of the posterior probability of the true model, $n = 100$

k	4		10		20		40	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	0.6993	0.1616	0.4016	0.1263	0.1745	0.0772	0.0111	0.0113
b	0.6562	0.1570	0.5970	0.1908	0.0007	0.0020	0.0000	0.0000
c	0.6408	0.1648	0.3034	0.1042	0.1116	0.0516	0.0097	0.0086
d	0.4599	0.1301	0.1589	0.0430	0.0115	0.0048	0.0000	0.0000
e	0.7269	0.1210	0.1947	0.1103	0.0000	0.0001	0.0000	0.0000
f	0.6971	0.1618	0.3986	0.1256	0.1722	0.0761	0.0107	0.0107
g	0.8036	0.1059	0.0604	0.0439	0.0000	0.0000	0.0000	0.0000
h	0.5503	0.1274	0.1676	0.0421	0.0035	0.0015	0.0000	0.0000
i	0.5084	0.1442	0.4015	0.1262	0.3145	0.1352	0.0949	0.1074

Table 8
Model 3: Means and Stds of the posterior probability of the true model, $n = 1000$

k	4		10		20		40	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	0.8794	0.0857	0.7029	0.1483	0.5349	0.1272	0.2366	0.1699
b	0.8560	0.0991	0.9515	0.0690	1.0000	0.0000	0.6499	0.4769
c	0.8452	0.1013	0.6081	0.1509	0.3950	0.1188	0.1853	0.0788
d	0.5896	0.1451	0.2780	0.0918	0.1208	0.0325	0.0459	0.0054
e	0.9979	0.0023	1.0000	0.0000	0.9395	0.2374	0.0000	0.0000
f	0.8110	0.1161	0.5744	0.1495	0.3755	0.1127	0.1972	0.0685
g	1.0000	0.0000	1.0000	0.0001	0.0015	0.0036	0.0000	0.0000
h	0.9958	0.0035	0.9357	0.0164	0.1731	0.1957	0.0222	0.0070
i	0.5150	0.1349	0.4174	0.1305	0.4030	0.1172	0.3170	0.1856

prior g). Interestingly, prior i, which is decreasing in k does not suffer from this, and performs quite well for $n = 100$. In fact, priors a, f and i all do remarkably well in identifying the true model from a very large model space on the basis of a mere 100 observations. The results for $k = 40$ appear less convincing, but we have to bear in mind that the posterior probability of the correct model multiplies the corresponding prior probability by more than 1×10^{10} with these priors.

When we consider Table 8, corresponding to $n = 1000$, we note that most priors benefit from the larger sample size. Only prior i leads to virtually the same posterior probabilities as with the smaller sample (except for large k). In line with the larger sample size n, the drop in posterior probability for the priors with g_{0j} increasing in k_j now tends to occur at a higher value of k than in Table 7.

5.3. Predictive inference

5.3.1. Results under Model 1

As discussed in Section 4.3 we shall condition our predictions on values of the regressors z_f. In all, we choose $q = 19$ different vectors for these regressors, and we shall focus especially on those vectors that lead to the minimum, median and maximum value for the mean of the sampling model. We shall denote these regressors as $z_{\min}$, z_{med} and $z_{\max}$, respectively. In our particular case, $z_{\min}$ will be more extreme than $z_{\max}$ for Model 1. For Model 3, both are roughly equally extreme.

Firstly, Table 9 presents the median of $LPS(z_f, y, Z)$ in (3.8), computed across the 100 samples (y, Z), conditionally upon the three vectors of regressors mentioned above. In interpreting these numbers, it is useful to recall that the theoretical minimum of the integral corresponding to LPS is 2.335, as explained

Table 9
Model 1: conditional medians of $LPS(z_f, y, Z)$

n	50			100			1000			10,000			100,000		
	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$
a	2.471	2.425	2.428	2.391	2.389	2.391	2.334	2.355	2.348	2.330	2.352	2.347	2.331	2.350	2.345
b	2.480	2.422	2.431	2.409	2.390	2.385	2.334	2.355	2.347	2.330	2.352	2.347	2.331	2.350	2.345
c	2.471	2.424	2.433	2.397	2.389	2.389	2.334	2.355	2.347	2.330	2.352	2.347	2.331	2.350	2.345
e	2.691	2.448	2.475	2.507	2.406	2.410	2.358	2.356	2.354	2.332	2.351	2.347	2.332	2.351	2.345
f	2.474	2.428	2.428	2.393	2.389	2.391	2.333	2.355	2.347	2.329	2.352	2.346	2.331	2.350	2.345
g	2.836	2.470	2.530	2.636	2.423	2.463	2.475	2.378	2.412	2.444	2.371	2.391	2.417	2.362	2.379
h	2.492	2.430	2.440	2.450	2.392	2.395	2.418	2.362	2.381	2.423	2.367	2.382	2.422	2.364	2.382
i	2.488	2.421	2.428	2.392	2.389	2.392	2.333	2.355	2.347	2.329	2.352	2.346	2.332	2.350	2.345

in Section 4.3. Of course, *LPS* in (3.8) is only a Monte Carlo approximation to this integral (based on a mere 100 draws), so this lower bound is not always strictly adhered to. The Monte Carlo (numerical) standard error corresponding to the numbers in Table 9 is roughly equal to 0.02 for $n = 50$, decreases to about 0.01 for $n = 100$ and then quickly settles down at about 0.007 as n becomes larger. Under priors a, b, c, f and i we are predicting the sampling density virtually exactly with samples of size $n = 1000$ or more. These same priors also lead to the best results for smaller n. Prior e performs worse for small samples, while priors g and h tend to be even further from the actual sampling density and do not lead to perfect prediction even with 100,000 observations.

In order to find out more about the differences between the predictive density in (3.6) and the sampling density in (3.3), we can overplot both densities for the three values of $z_{\min}, z_{\text{med}}$ and $z_{\max}$. Fig. 1 displays this comparison for different values of n and the predictives for 25 of the 100 generated samples (to avoid

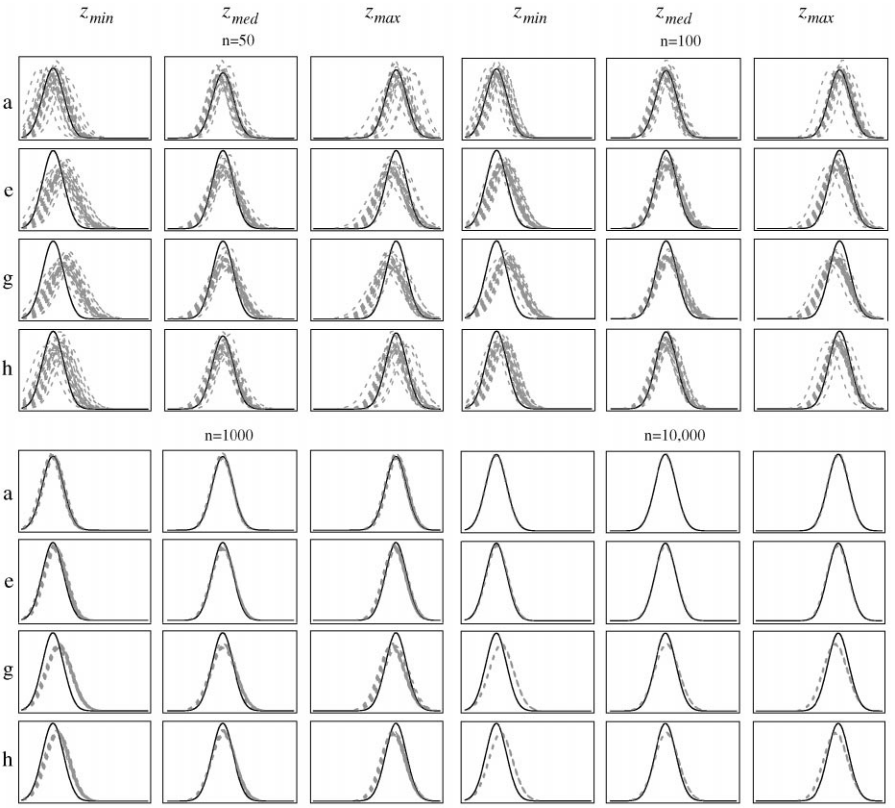

Fig. 1. Model 1: Predictive densities, $n = 50$, 100, 1000 and 10,000.

cluttering the graphs). The drawn line corresponds to the actual sampling density. Since the predictives from priors a–c, f and i are very close for all sample sizes, we shall only present the graphs for priors a, e, g and h. It is clear that for $n = 50$ substantial uncertainty remains about the predictive distribution: different samples can lead to rather different predictives. They are, however fairly well calibrated in that they tend to lie on both sides of the actual sampling density for priors a–c, f and i and there is no clear tendency towards a different degree of concentration. These are exactly the priors for which g_{0j} takes on fairly small values (for prior i $g_{0j} = 0.0044$ and for the others it is in between 0.02 and 0.08). Priors e, g and h lead to much larger values for g_{0j} (in the range 0.16 to 0.41) and show a clear tendency for the predictive densities to be somewhat biased towards the median when conditioning on $z_{\min}$ and $z_{\max}$. In addition, these priors induce predictives that are, on average, less concentrated than the sampling density, even for very large sample sizes. This behaviour can easily be understood once we realize that the location of (3.6) (the posterior mean of $\alpha + z'_{f,j}\beta_j$) is clearly shrunk more towards the sample mean $\boldsymbol{y}$ as g_{0j} becomes larger. This is, of course, in accordance with the zero prior mean for β_j and the g-prior structure in (2.15). In addition, predictive precision decreases with g_{0j}, which explains the systematic excess spread of the predictives with respect to the sampling density for priors e, g and h. As sample size increases, the predictive distributions get closer and closer to the actual sampling distribution, and for $n = 1000$ or larger the effect of shrinkage due to g_{0j} has become negligible for prior e (g_{0j} is then equal to 0.07 for $k_j = 5$) whereas it persists for priors g and h even with 100,000 observations (where g_{0j} takes the value 0.14 and 0.16, respectively). The graph for the latter case is not presented in Fig. 1, but it is very close to that with $n = 10{,}000$.

We can also compare overall predictive performance, through considering $LPS(z_f, y, Z)$ for the 19 different values of z_f and the 100 samples of (y, Z). This leads to the results presented in Table 10, where the medians (computed across the 1900 sample-z_f combinations) are recorded for the different priors and sample sizes. Monte Carlo standard errors corresponding to the entries in Table 10 are approximately 0.005 for $n = 50$, about 0.0025 for $n = 100$ and decrease to around 0.0015 for larger n. Clearly, whereas all priors except for priors e and g lead to comparable predictive behaviour for very small n, the fact that g_{0j} is large and constant in n makes prior h lose ground with respect to the other priors as n increases. Prior g always performs worse than priors a through f and prior i. Note that the latter priors lead to median LPS values that are roughly equal to the theoretical minimum of 2.335, implying perfectly accurate prediction, for $n \geqslant 10{,}000$ (which may, of course, be quite a large sample size for some applications). Priors a–c, f and i perform best overall, and are already quite close to perfect prediction for $n = 1000$.

In addition, we can compare the percentiles of the sampling distribution and the predictive in (3.6). We have computed the predictive percentiles corresponding

Table 10
Model 1: medians of $LPS(z_f, y, Z)$

n	50	100	1000	10,000	100,000
a	2.427	2.382	2.339	2.334	2.333
b	2.427	2.383	2.339	2.334	2.333
c	2.424	2.381	2.339	2.334	2.333
e	2.473	2.416	2.347	2.336	2.334
f	2.428	2.382	2.339	2.334	2.333
g	2.502	2.452	2.393	2.374	2.363
h	2.433	2.452	2.369	2.367	2.366
i	2.428	2.382	2.339	2.334	2.334

to the 1, 5, 25, 50, 75, 95, and 99 sampling percentile. The quartiles of these numbers, calculated over all 1900 sample-z_f combinations, confirm that priors a–c, f and i lead to better predictions for small sample sizes, where the predictives from e, g and h are too spread out. Starting at $n = 1000$, prior e predicts well, whereas the inaccurate predictions with priors g and h persist even for very large sample sizes. For space considerations, these results are not presented here, but are obtainable upon request from the authors. In addition, Fig. 1 shows that most of the spread in the percentiles for priors g and h with $n \geqslant 10{,}000$ is due to the bias towards the median (shrinkage).

5.3.2. Results under Model 2

As mentioned in Section 5.2.2, it is hard to correctly identify the null model when we generate the data from such a model. On the other hand, prediction seems much easier than model choice. This can immediately be deduced from predictive percentiles (not reported, but obtainable upon request). The incorrectly chosen models are such that they do not lead our predictions (averaged over all the chosen models as in (3.6)) far astray. Even with just 50 observations median predictions are virtually exact, and the spread around these values is relatively small. Moreover, this behaviour is encountered for all priors. When sample size is up to 1000, prediction is near perfect for all priors. For this model the issue of shrinkage is, of course, less problematic.

5.3.3. Results under Model 3

Table 11 presents the medians of the conditional $LPS(z_f, y, Z)$ as in (3.8), computed with $v = 100$ out-of-sample observations. The theoretical minimum corresponding to perfect prediction is 2.112, and Monte Carlo standard errors for the numbers in Table 11 are of the order 0.008 for $k = 4$ and vary from 0.01 to 0.04 for $k = 40$. The most obvious finding from Table 11 is the poor performance of priors e, g and h, which correspond to the largest values of g_{0j}. These priors, as well as prior b, do progressively worse as k increases since g_{0j} is

Table 11
Model 3: conditional medians of $LPS(z_f, y, Z)$, $n = 100$

k	4			10			20			40		
	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$	$z_{\min}$	z_{med}	$z_{\max}$
a	2.151	2.110	2.138	2.154	2.139	2.168	2.220	2.218	2.224	2.555	2.453	2.490
b	2.153	2.111	2.133	2.200	2.164	2.202	2.646	2.529	2.588	3.365	3.161	3.226
c	2.153	2.111	2.136	2.160	2.142	2.168	2.227	2.223	2.230	2.574	2.467	2.492
d	2.173	2.113	2.151	2.244	2.195	2.287	2.592	2.502	2.562	3.215	3.076	3.120
e	2.192	2.120	2.171	2.449	2.293	2.584	3.089	2.802	3.003	3.802	3.422	3.764
f	2.151	2.110	2.138	2.154	2.140	2.168	2.221	2.218	2.224	2.557	2.461	2.497
g	2.240	2.129	2.221	2.615	2.395	2.883	3.328	2.916	3.237	4.034	3.552	4.024
h	2.193	2.119	2.169	2.341	2.248	2.434	2.832	2.648	2.769	3.434	3.251	3.409
i	2.155	2.113	2.133	2.154	2.139	2.168	2.212	2.219	2.232	2.472	2.401	2.471

an increasing function of k_j. The best forecasting performance is observed for priors a, c, f and i, which all do a very good job, even in the context of a very large model set.

5.4. Recommendations

On the basis of the posterior and predictive results mentioned above, we can offer the following advice to practitioners, although we stress that it is probably impossible to recommend a choice for g_{0j} that performs optimally in all situations. Nevertheless, it appears that a strategy using prior i for the smallest values of n, in particular where $n \leqslant k^2$, and prior a for those cases where $n > k^2$ would do very well in most situations. It would actually lead to the best or close to the best results in all comparisons made above. The only substantial improvement to that strategy would be to choose prior e instead of prior a for Model 1. Whereas prediction would be slightly less good, posterior probabilities of the true model would be quite a bit higher. A similar, but less clear-cut, trade-off can be observed for Model 3. However, this strategy appears more risky, as it would lead to a dramatic fall in performance for Model 2 on the same criterion (see Table 5). Of course, Model 2 might be considered a very unusual situation, so that some practitioners may well adopt this more risky strategy, especially if they are more interested in posterior results, rather than in predicting. As a general recommendation, however, we would propose the 'safe' strategy.

Note that the latter implies choosing the prior rule with the smallest value of g_{0j}, i.e., it amounts to

$$g_{0j} = \frac{1}{\max\{n, k^2\}} \tag{4.1}$$

and is quite unlikely to lead us far astray, since g_{0j} will never take large values in situations of practical relevance. This prior will combine the consistency properties of prior a with the impressive small sample performance of prior i.

6. An empirical example: Crime data

The literature on the economics of crime has been critically influenced by the seminal work of Becker (1968) and the empirical analysis of Ehrlich (1973, 1975). The underlying idea is that criminal activities are the outcome of some rational economic decision process, and, as a result, the probability of punishment should act as a deterrent. Raftery et al. (1997) have used the Ehrlich data set corrected by Vandaele (1978). These are aggregate data for 47 U.S. states in 1960, which will be used here as well.

The single-equation cross-section model used here is not meant to be a serious attempt at an empirical study of these phenomena. For example, the model does not address the important issues of simultaneity and unobserved heterogeneity, as stressed in Cornwell and Trumbull (1994), and the data are at state level, rather than individual level, but we shall use it mainly for comparison with the results in Raftery et al. (1997), who also treat it as merely an illustrative example.

We shall, thus, consider a linear regression model as in (1.1), where the dependent variable, y, groups observations on the crime rate, and the 15 regressors in Z are given by: percentage of males aged 14–24, dummy for southern state, mean years of schooling, police expenditure in 1960, police expenditure in 1959, labour force participation rate, number of males per 1000 females, state population, number of nonwhites per 1000 people, unemployment rate of urban males aged 14–24, unemployment rate of urban males aged 25–39, wealth, income inequality, probability of imprisonment, and average time served in state prisons. All variables except for the southern dummy are transformed to logarithms.

In line with the recommendations from Section 5, we shall use prior i for this very small sample ($n = 47$). We run the MC^3 chain to produce 100,000 draws after a burn-in of 25,000. This is more than enough to achieve convergence, as is evidenced by the near perfect correlation (0.9930) between the model probabilities based on the actual Bayes factors computed as in (2.16) and the relative frequencies of model visits. All results will be based on the actual Bayes factors of the models visited (as explained in Section 5.1). In all, 2465 different models were visited, and the best 10% of those models account for 80.1% of the posterior model probability. Thus, posterior mass is not highly concentrated on just a few models. As proposed in George and McCulloch (1997, Section 4.5), we can consistently estimate the total posterior probability of all visited models by contrasting the sum of the marginal likelihoods and the relative frequencies of visits for a fixed subset of models. Taking for that subset the ten best models of

Table 12
Crime data: models with more than 2% posterior probability

	Prob. (%)	Included regressors								
1	3.61	1	3	4				11	13	14
2	3.48	1	3	4			9	11	13	14
3	2.33	1	3		5			11	13	14
4	2.31	1	3	4				11	13	
5	2.29	1	3	4					13	14
6	2.20	1	3		5		9	11	13	14
7	2.05		3	4		8	9		13	14
8	2.02	1	3	4			9		13	14

Raftery et al. (1997, Table 3), we estimate the total model probability covered by the chain to be 99.3%, thus corroborating our earlier conclusion of convergence. Note that this run takes a mere 34 seconds on a 200 MHz PPC603ev-based Macintosh 3400c laptop computer.

Table 12 presents the 8 models that receive over 2% posterior probability. Seven of these models are among the ten best models of Raftery et al. (1997). In general, model probabilities are roughly similar, even though our prior is quite different from the one proposed in Raftery et al. (1997). In particular, we only require the user to choose the function g_{0j}, and choosing it in accordance with our findings in Section 4 leads to results that are reasonably close to those with the rather laboriously elicited prior of Raftery et al. (1997). The second best model in Table 12 is the one that is chosen by the Efroymson stepwise regression method, as explained in Raftery et al. (1997), and is the model with highest posterior probability in the latter paper. If we use prior a in our framework, we also get this as the model with highest posterior probability. Generally, prior a leads to a more diffuse posterior model probability and larger models.

Posterior probabilities of including each of the regressors are given in Table 13, which clearly indicates that schooling and income inequality are virtually always included, while the percentage of males aged 14–24 and the probability of imprisonment are also typically part of the relevant models. Overall, Table 13 roughly agrees with Table 4 in Raftery et al. (1997). The deterrence variables are probability of imprisonment and average time served in prisons. These variables are of particular interest for the economic theory of crime, and their marginal posterior density functions (averaging over models with posterior probabilities) are given in Fig. 2. The coefficients of these regressors can be interpreted as elasticities. The gauge on top indicates (in black) the posterior probability of inclusion. Thus, the full posterior distribution of these coefficients contains a point mass ('spike') at zero with a probability proportional to the grey part of the gauge (since the scaling of this spike cannot be related to the continuous

Table 13
Crime data: posterior probabilities of including each regressor

	Regressor	Prob. (%)
1	Percentage of males age 14–24	75.8
2	Indicator variable for southern state	14.1
3	Mean years of schooling	95.5
4	Police expenditure in 1960	65.8
5	Police expenditure in 1959	38.1
6	Labor force participation rate	7.5
7	Number of males per 1000 females	8.5
8	State population	22.2
9	Number of nonwhites per 1000 people	50.7
10	Unemployment rate for urban males, age 14–24	10.6
11	Unemployment rate for urban males, age 25–39	45.1
12	Wealth	17.5
13	Income inequality	99.8
14	Probability of imprisonment	78.9
15	Average time served in prisons	18.0

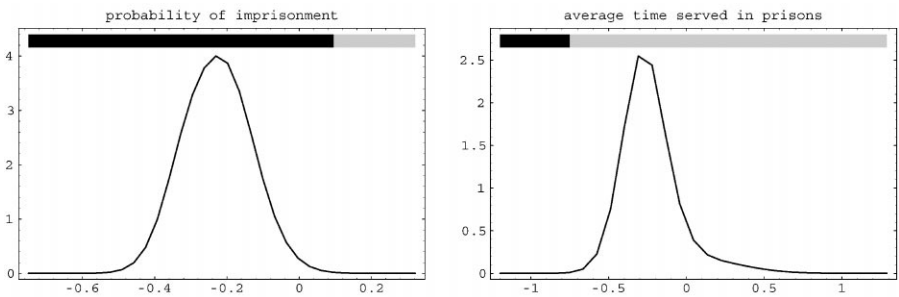

Fig. 2. Posterior density functions: regressors 14 and 15.

part, we have opted for the current presentation). The probability of imprisonment seems to have a moderately negative influence, as expected. The average time served in prisons, however, only has a posterior probability of inclusion of under 20% (see also Table 13).[5]

[5] Fig. 2 again illustrates the fact that we allow for the formal exclusion of a regressor through a point mass at zero, and we do not follow the strategy of identifying the mass around zero in a continuous distribution for a regression coefficient with the exclusion probability of the corresponding regressor. Instead, we conduct inference on the exclusion probabilities through formal posterior odds. Given the fact that Bayes factors tend to favour parsimony — see e.g., Smith and Spiegelhalter (1980) — it is not surprising that, while many models exclude variable 15, the ones that do include it assign a continuous distribution to β_{15} that has little mass around zero.

As two referees pointed out to us, police expenditure in 1959 and that in 1960 constitute two highly collinear regressors. Note from Table 12 that the models with high posterior probability include either one or the other, but never both (this holds true for the 69 best models with posterior probability over 0.27%). Interestingly, Raftery et al. (1997, Table 3) report a model with both variables as one of their 10 best models. Clearly, this level of detail is not obvious from the posterior probabilities of inclusion presented in Table 13, which mainly tells us that regressors 3 and 13 are virtually always included, whereas perhaps only regressors 6 and 7 could be ignored without appreciable empirical consequences.

If we split the data randomly into observations used for estimation (with probability 0.75) and observations to be predicted, we can construct the predictive Q–Q plots in Fig. 3. These are what Raftery et al. (1997) call 'calibration plots'. For a given prediction sample we record in which percentile of the predictive distribution (using the corresponding values of the regressors) the actual observations fall. Plotting predictive quantiles against empirical ones thus obtained leads to this Q–Q plot. The closer the plot is to the 45° line, the better the model (estimated on the basis of the inference sample) does in predicting the data in the prediction sample. Fig. 3 contains Q–Q plots for 10 different random partitions of the data, and two different models in each case: the single model with the highest posterior probability (which generally changes in each partition) and the averaged model through BMA. We note that the best single model (dotted lines) tends to predict worse than the predictive in (3.6) resulting from BMA (drawn lines). We tried a total of 50 different data partitions and found that BMA outperformed the best single model 14 times and was beaten 8 times. In the other 28 cases performance was about the same. This is suggestive, although not conclusive, evidence of the predictive superiority of BMA. Of course, in a model selection strategy, inference could simply be based on the best model (as identified in Table 12).

7. Conclusions

We consider the normal linear regression model with uncertainty regarding the choice of regressors. The prior structure we have proposed in Section 3 leads to a valid interpretation of the posterior distribution as a conditional and only requires the choice of one scalar hyperparameter, called g_{0j}. We make g_{0j} a possible function of the sample size, n, the number of regressors in the model under consideration, k_j, and the total number of available regressors, k. Theoretical results on consistency (in the sense of correctly identifying the model that generated the data if that model is contained in model space) suggest making g_{0j} a decreasing function of sample size n. In addition, empirical results on posterior model choice and predictive performance in the context of an

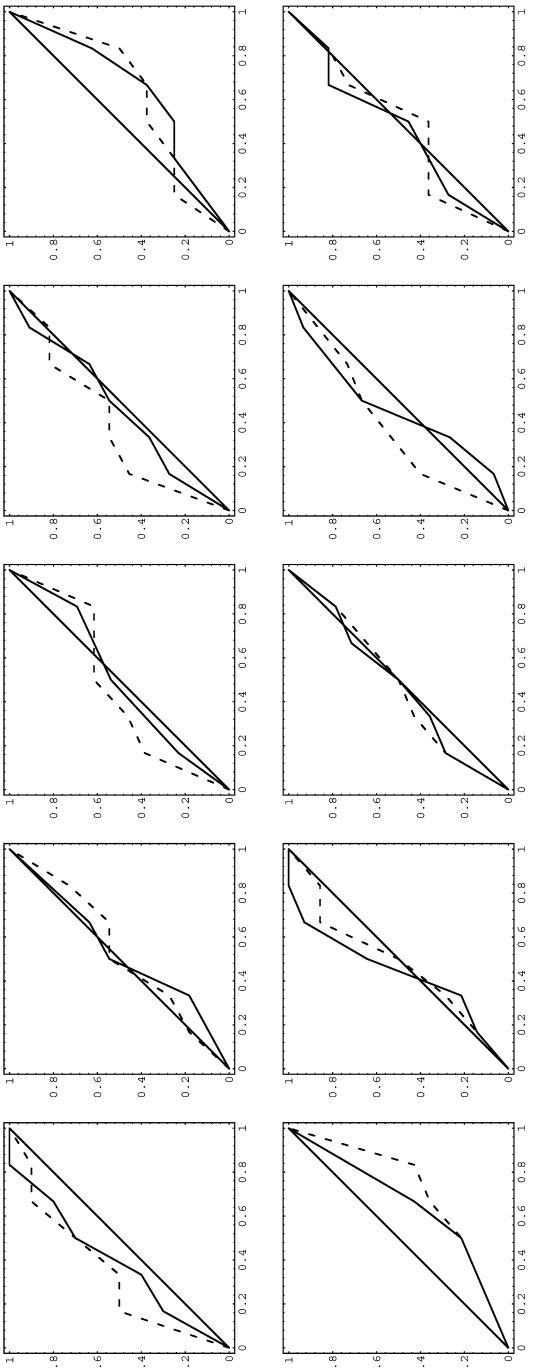

Fig. 3. Q–Q plots with 75–25% sample split.

extensive simulation study indicate that the following strategy is a reasonable choice:

- prior i, where $g_{0j} = 1/k^2$, for $n \leqslant k^2$,
- prior a, where $g_{0j} = 1/n$, for $n > k^2$,

which leads to consistent Bayes factors and corresponds to choosing g_{0j} as in (4.1).

Thus, we would recommend the prior structure introduced here, together with these choices of g_{0j} when faced with uncertain variable selection in linear regression models, whenever substantial prior information is lacking or a default analysis is the aim.

Acknowledgements

We thank Arnold Zellner, Dennis Lindley and two anonymous referees for their useful suggestions. Carmen Fernández gratefully acknowledges financial support from a Training and Mobility of Researchers grant awarded by the European Commission (ERBFMBICT # 961021). Carmen Fernández and Mark Steel were affiliated to CentER and the Department of Econometrics, Tilburg University, The Netherlands, and Eduardo Ley was at FEDEA, Madrid, Spain during the early stages of the work on this paper. Some of this research was done when Carmen Fernández was at the Department of Mathematics, University of Bristol, and Mark Steel at the Department of Economics, University of Edinburgh.

Appendix A. Asymptotic properties and the choice of g_{0j}

We focus on the prior given through (2.11), (2.14) and (2.15), which leads to the expression in (2.16) for the Bayes factor. Although the importance of asymptotic results for statistics is certainly a contentious issue (see Lindsey, 1999, for a stimulating discussion), we feel they have some role to play in assessing the general properties of statistical procedures. This appendix is included for the benefit of the interested reader.

Throughout this appendix we assume that the sample y is generated by model $M_s \in \mathscr{M}$ with parameter values α, β_s and σ, i.e.,

$$y = \alpha \iota_n + Z_s \beta_s + \sigma \varepsilon. \tag{A.1}$$

First, we examine consistency in the sense that

$$\operatorname{plim}_{n\to\infty} P(M_s|y) = 1 \quad \text{and} \quad \operatorname{plim}_{n\to\infty} P(M_j|y) = 0 \quad \text{for all } M_j \neq M_s, \tag{A.2}$$

where the probability limit is taken with respect to the true sampling distribution described in (A.1). By (1.7), as long as the prior (1.5) on the model space does not depend on sample size, we simply need to check that the Bayes factor for model M_j versus model M_s, B_{js}, converges in probability to zero for any model M_j other than M_s. The reference posterior odds proposed in Bernardo (1980) and Pericchi (1984) rely on making prior model probabilities depend on the expected gain in information from the sample. As explained in these papers, such procedures will generally not lead to consistency in the sense of (A.2).

Although we shall focus on the case of improper priors on α and σ, thus leading to the expression for B_{js} in (2.16), it is immediate to see that the same results apply to the Bayes factor in (2.9) (which corresponds to proper priors on both α and σ) and to the Bayes factor in (2.12) (where we are still proper on α).

Before examining consistency of the Bayes factors, we need to establish some preliminary results. Some of these results are not new, whereas others are easy to derive.

Lemma A.1. Under the sampling model M_s in (A.1),

(i) If M_s is nested within or is equal to model M_j,

$$\operatorname{plim}_{n\to\infty} \frac{y' M_{X_j} y}{n} = \sigma^2. \tag{A.3}$$

(ii) Under the assumption that for any model M_j that does not nest M_s,

$$\lim_{n\to\infty} \frac{(\alpha,\ \beta_s')X_s' M_{X_j} X_s(\alpha,\ \beta_s')'}{n} = b_j \in (0, \infty), \tag{A.4}$$

we obtain

$$\operatorname{plim}_{n\to\infty} \frac{y' M_{X_j} y}{n} = \sigma^2 + b_j. \tag{A.5}$$

In the sequel, we shall assume (A.4) to hold. We now examine two different functional choices for g_{0j}. We first consider dependence on the sample size n and, possibly, on the number of regressors k_j, and afterwards we suppress the dependence on n. Some of our choices for g_{0j} will also depend on k, the total number of regressors in the information set, but our notation will not reflect this, since k is assumed fixed throughout the appendix.

A.1. Results under $g_{0j} = w_1(k_j)/w_2(n)$ with $\lim_{n\to\infty} w_2(n) = \infty$

Through (2.15), the prior precision is a fraction g_{0j} of the sample precision. It seems logical to assume that the relative precision of the prior vanishes as n goes to infinity. In addition, we let g_{0j} depend on a function w_1 of k_j.

Theorem A.1. Consider the Bayesian model given by (1.1), together with the prior densities in (2.11), (2.14), (2.15) and any prior on the model space $\mathcal{M}$ in (1.5). We assume that g_{0j} in (2.15) takes the form

$$g_{0j} = \frac{w_1(k_j)}{w_2(n)} \quad \text{with} \quad \lim_{n\to\infty} w_2(n) = \infty\,. \tag{A.6}$$

Then, under the assumption that there is a true model M_s in $\mathcal{M}$ that generates the data, the condition

$$\lim_{n\to\infty} \frac{w_2'(n)}{w_2(n)} = 0, \tag{A.7}$$

together with either

$$\lim_{n\to\infty} \frac{n}{w_2(n)} \in [0, \infty) \tag{A.8}$$

or

$$w_1(\cdot) \quad \text{is a non-decreasing function,} \tag{A.9}$$

ensures that the posterior distribution of the models is consistent in the sense defined in (A.2).

Proof. Denoting by C_{js} the product of the first two factors in (2.16), we have that

$$C_{js} = \left(\frac{w_1(k_j)}{g_{0j}+1}\right)^{k_j/2} \left(\frac{g_{0s}+1}{w_1(k_s)}\right)^{k_s/2} w_2(n)^{(k_s-k_j)/2}, \tag{A.10}$$

and thus

$$\lim_{n\to\infty} C_{js} = \begin{cases} 0 & \text{if } k_j > k_s, \\ 1 & \text{if } k_j = k_s, \\ \infty & \text{if } k_j < k_s. \end{cases} \tag{A.11}$$

On the other hand, the limiting behaviour of the last factor in (2.16) depends on whether M_s is nested within M_j. We therefore consider the three following situations:

A.1.1. M_s is not nested within M_j and $k_j \geqslant k_s$

Denoting by D_{js} the last factor in (2.16) and applying (A.5) we obtain

$$\operatorname*{plim}_{n\to\infty} D_{js} = \lim_{n\to\infty} \left(\frac{\sigma^2}{\sigma^2 + b_j}\right)^{(n-1)/2} = 0. \tag{A.12}$$

Combining the latter result with (*A*.11) directly leads to a zero limit for B_{js}.

A.1.2. M_s is not nested within M_j and $k_j < k_s$

In this case, combining (A.11) with (A.12) no longer leads directly to the limit of B_{js}. A natural *sufficient condition* leading to a zero limit for B_{js} is

$$\lim_{n\to\infty} \frac{w_2'(n)}{w_2(n)} = 0, \tag{A.13}$$

which ensures that $w_2(n)^{(k_s-k_j)/(n-1)}$ converges to unity.

A.1.3. M_s is nested within M_j

Since in this case $k_j > k_s$, we know from (A.11) that C_{js} converges to zero. However, the limit of D_{js} is now difficult to assess. Here we shall present *sufficient conditions* for a zero limit of B_{js}.

Rewriting D_{js} as

$$D_{js} = \left(\frac{y'M_{X_s}y}{y'M_{X_j}y}\right)^{(n-1)/2} \times\left(1+\frac{w_2(n)\{w_1(k_s)(A_s-1)-w_1(k_j)(A_j-1)\}+w_1(k_s)w_1(k_j)(A_s-A_j)}{\{w_2(n)+w_1(k_s)\}\{w_2(n)+w_1(k_j)A_j\}}\right)^{(n-1)/2}, \tag{A.14}$$

where $A_s = (y-\bar{y}\iota_n)'(y-\bar{y}\iota_n)/y'M_{X_s}y$ and $A_j = (y-\bar{y}\iota_n)'(y-\bar{y}\iota_n)/y'M_{X_j}y$, it is immediate that the first factor converges in distribution to $\exp(S/2)$, where S has a χ^2 distribution with k_j-k_s degrees of freedom. On the other hand, the condition

$$\lim_{n\to\infty} \frac{n}{w_2(n)} \in [0,\infty) \tag{A.15}$$

ensures a finite limit for the second factor. Alternatively, if

$$w_1(\cdot) \quad \text{is a nondecreasing function,} \tag{A.16}$$

the second factor is (A.14) is smaller than one. Thus, using the fact that C_{js} converges to zero, (A.15) and (A.16) each provide a sufficient condition for a zero limit of B_{js}.

On the basis of prior ideas about fit, Poirier (1996) suggests taking $w_2(n) = n$, which satisfies (A.7) and (A.8) and thus leads to consistent Bayes factors. This and other choices are discussed in the main text.

The proof of Theorem A.1 never makes use of the Normality assumption for the error distribution of the 'true' model in (A.1), and, thus, our findings immediately generalize to the case where the components of ε in (A.1) are i.i.d. following any regular distribution (i.e., leading to asymptotic normality) with finite variance. Therefore, even if the true model does not possess a normally distributed error term, the posterior distribution derived on the basis of the

models with normality assumed [leading to the Bayes factor in (2.16)] is still consistent, in the sense of asymptotically selecting the true subset of regressors, under the sufficient conditions for g_{0j} stated in Theorem A.1. This implies that we can always make the convenient assumption of normality to asymptotically select the correct set of regressors. In some sense, this offers a counterpart to the classical result for testing nested models, where the likelihood ratio, Wald and Rao (or Lagrange multiplier) statistics derived under the assumption of normality keep the same asymptotic distribution (a χ^2) even if the error term is non-normal — see, e.g., Amemiya (1985, p. 144).

A.2. Results with $g_{0j} = w(k_j)$

We now examine the situation where g_{0j} is no longer a function of the sample size n. Therefore, consistency is entirely driven by the last factor of B_{js} in (2.16), which we denote by D_{js}.

It is immediately clear that in this situation we do not have consistency: When the data generating model, M_s, is the model with just the intercept, $D_{js} \geqslant 1$ regardless of the data [since the numerator in the last factor of (2.16) is then $(y - \bar{y}\iota_n)'(y - \bar{y}\iota_n)$, which is always bigger than or equal to the denominator]. Thus, $P(M_s|y)$ can not converge to one as n tends to infinity, precluding consistency.

Even though we do not have consistency, let us examine the asymptotic behaviour of D_{js} for the case where M_s contains some regressors other than the intercept (i.e., $k_s \geqslant 1$).

When M_s is nested within M_j, we have that $y'M_{X_s}y \geqslant y'M_{X_j}y$. As a consequence, having a zero limit for the Bayes factor requires that $w(\cdot)$ be an increasing function, since otherwise $D_{js} \geqslant 1$. Provided that $w(\cdot)$ has this property, we obtain

$$\operatorname*{plim}_{n\to\infty} D_{js} = \lim_{n\to\infty} \left(\frac{\sigma^2 + (g_{0s}/(g_{0s}+1))b}{\sigma^2 + (g_{0j}/(g_{0j}+1))b} \right)^{(n-1)/2} = 0, \tag{A.17}$$

where b denotes the value b_j in (A.4) corresponding to $X_j = \iota_n$. This immediately leads to:

$$\operatorname*{plim}_{n\to\infty} D_{js} = 0 \quad \text{if and only if } w(\cdot) \text{ is an increasing function.} \tag{A.18}$$

The situation becomes less clear-cut when M_s is not nested within M_j. Some results can be obtained upon request from the authors.

A.3. Relationship to information criteria

A number of information criteria have traditionally been used for classical model selection purposes, especially in the area of time series analysis. In this

subsection, we shall establish asymptotic links between the Bayes factors corresponding to Section A.1 and two consistent information criteria: the Schwarz (or Bayes information) criterion as derived in Schwarz (1978) and the criterion of Hannan and Quinn (1979). If we wish to compare two models as in (1.1), say M_j versus M_s, these criteria take the form

$$S_{js} = \frac{n}{2} \ln\left(\frac{y'M_{X_s}y}{y'M_{X_j}y}\right) + \frac{k_s - k_j}{2} \ln(n), \tag{A.19}$$

$$HQ_{js} = \frac{n}{2} \ln\left(\frac{y'M_{X_s}y}{y'M_{X_j}y}\right) + \frac{k_s - k_j}{2} C_{HQ} \ln\ln(n). \tag{A.20}$$

Hannan and Quinn (1979) prove strong consistency for both criteria provided $C_{HQ} > 2$.

The asymptotic behaviour of the Bayes factor in (2.16), made consistent by choosing g_{0j} as in Theorem A.1 can be characterized by the following result:

Theorem A.2. Consider the Bayesian model described in Theorem A.1, with g_{0j} satisfying (A.7) together with either (A.8) or (A.9). Then the Bayes factor in (2.16) satisfies

$$\operatorname{p\,lim} \frac{\ln B_{js}}{(n/2)\ln(y'M_{X_s}y/y'M_{X_j}y) + ((k_s - k_j)/2)\ln w_2(n)} = 1, \tag{A.21}$$

where the probability limit is taken with respect to the model M_s as described in (A.1).

Thus, different choices of the function $w_2(n)$ will influence the asymptotic behaviour of the logarithm of the Bayes factor. In particular, let us consider the choices of $w_2(n)$ that induce a relationship with the two information criteria mentioned above.

Corollary A.1. If in Theorem A.2 we choose $w_2(n) = n$, we obtain

$$\operatorname{p\,lim} \frac{\ln B_{js}}{S_{js}} = 1, \tag{A.22}$$

whereas choosing $w_2(n) = \{\ln(n)\}^{C_{HQ}}$ and $w_1(\cdot)$ non-decreasing, leads to

$$\operatorname{p\,lim} \frac{\ln B_{js}}{HQ_{js}} = 1. \tag{A.23}$$

From these results we see that $\ln B_{js}$ behaves like these consistent criteria if we choose $w_2(n)$ appropriately. Note that the second choice of $w_2(n)$ in Corollary A.1 does not satisfy (A.8), which is why we impose that $w_1(\cdot)$ fulfills (A.9). Kass

and Wasserman (1995) study the relationship between the Schwarz criterion and Bayes factors using 'unit information priors' for testing nested hypotheses, and provide the order of the approximation under certain regularity conditions. See also George and Foster (1997) for Bayesian calibration of information criteria.

As a final note, it is again worth mentioning that Theorem A.2 also holds if the error terms in (A.1) follow a non-normal distribution.

References

Akaike, H., 1981. Likelihood of a model and information criteria. Journal of Econometrics 16, 3–14.

Amemiya, T., 1985. Advanced Econometrics. Blackwell, Oxford.

Atkinson, A.C., 1981. Likelihood ratios, posterior odds and information criteria. Journal of Econometrics 16, 15–20.

Bauwens, L., 1991. The pathology of the natural conjugate prior density in the regression model. Annales d'Economie et de Statistique 23, 49–64.

Becker, G.S., 1968. Crime and punishment: an economic approach. Journal of Political Economy 76, 169–217.

Berger, J.O., Pericchi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91, 109–122.

Bernardo, J.M., 1979. Expected information as expected utility. The Annals of Statistics 7, 686–690.

Bernardo, J.M., 1980. A Bayesian analysis of classical hypothesis testing (with discussion). In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics. University Press, Valencia, pp. 605–618.

Box, G.E.P., 1980. Sampling and Bayes' inference in scientific modelling and robustness (with discussion). Journal of the Royal Statistical Society, Series A 143, 383–430.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49, 327–335.

Chipman, H., 1996. Bayesian variable selection with related predictors. Canadian Journal of Statistics 24, 17–36.

Chow, G.C., 1981. A comparison of the information and posterior probability criteria for model selection. Journal of Econometrics 16, 21–33.

Clyde, M., Desimone, H., Parmigiani, G., 1996. Prediction via orthogonalized model mixing. Journal of the American Statistical Association 91, 1197–1208.

Cornwell, C., Trumbull, W.N., 1994. Estimating the economic model of crime with panel data. Review of Economics and Statistics 76, 360–366.

Dawid, A.P., 1984. Statistical theory: the prequential approach. Journal of the Royal Statistical Society, Series A 147, 278–292.

Dawid, A.P., 1986. Probability forecasting. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), Encyclopedia of Statistical Sciences, Vol. 7. Wiley, New York, pp. 210–218.

Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society, Series B 57, 45–97.

Ehrlich, I., 1973. Participation in illegitimate activities: a theoretical and empirical investigation. Journal of Political Economy 81, 521–567.

Ehrlich, I., 1975. The deterrent effect of capital punishment: a question of life and death. American Economic Review 65, 397–417.

Foster, D.P., George, E.I., 1994. The risk inflation criterion for multiple regression. The Annals of Statistics 22, 1947–1975.

Freedman, D.A., 1983. A note on screening regressions. The American Statistician 37, 152–155.

Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. Journal of the American Statistical Association 74, 153–160.
Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society, Series B 56, 501–514.
George, E.I., 1999. Bayesian model selection, Encyclopedia of Statistical Sciences Update, Vol. 3. In: Kotz, S., Read, C., Banks, D.L. (Eds.), Wiley, New York.
George, E.I., Foster, D.P., 1997. Calibration and empirical Bayes variable selection. Mimeo, University of Texas, Austin.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881–889.
George, E.I., McCulloch, R.E., 1997. Approaches for Bayesian variable selection. Statistica Sinica 7, 339–373.
Geweke, J., 1996. Variable selection and model comparison in regression. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5. Oxford University Press, Oxford, pp. 609–620.
Good, I.J., 1952. Rational decisions. Journal of the Royal Statistical Society, Series B 14, 107–114.
Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.
Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B 41, 190–195.
Hoeting, J.A., Raftery, A.E., Madigan, D., 1995. Simultaneous variable and transformation selection in linear regression. Technical Report 9506, Statistics Department, Colorado State University.
Hoeting, J.A., Raftery, A.E., Madigan, D., 1996. A method for simultaneous variable selection and outlier identification in linear regression. Computational Statistics and Data Analysis 22, 251–271.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.
Kass, R.E., 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.
Laud, P.W., Ibrahim, J.G., 1995. Predictive model selection. Journal of the Royal Statistical Society, Series B 57, 247–262.
Laud, P.W., Ibrahim, J.G., 1996. Predictive specification of prior model probabilities in variable selection. Biometrika 83, 267–274.
Leamer, E.E., 1978. Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley, New York.
Lee, H., 1996. Model selection for consumer loan application data, Mimeo, Technical Report 650, Statistics Department, Carnegie Mellon University.
Lindsey, J.K., 1999. Some statistical heresies (with discussion). The Statistician 48, 1–40.
Madigan, D., Gavrin, J., Raftery, A.E., 1995. Eliciting prior information to enhance the predictive performance of Bayesian graphical models. Communications in Statistics, Theory and Methods 24, 2271–2292.
Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's Window. Journal of the American Statistical Association 89, 1535–1546.
Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. International Statistical Review 63, 215–232.
Min, C., Zellner, A., 1993. Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. Journal of Econometrics 56, 89–118.
Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression (with discussion). Journal of the American Statistical Association 83, 1023–1036.
O'Hagan, A., 1995. Fractional Bayes factors for model comparison (with discussion). Journal of the Royal Statistical Society, Series B 57, 99–138.

Osiewalski, J., Steel, M.F.J., 1993. Regression models under competing covariance structures: a Bayesian perspective. Annales d'Economie et de Statistique 32, 65–79.

Pericchi, L.R., 1984. An alternative to the standard Bayesian procedure for discrimination between Normal linear models. Biometrika 71, 575–586.

Phillips, P.C.B., 1995. Bayesian model selection and prediction with empirical applications (with discussion). Journal of Econometrics 69, 289–365.

Poirier, D., 1985. Bayesian hypothesis testing in linear models with continuously induced conjugate priors across hypotheses. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics 2. Elsevier, New York, pp. 711–722.

Poirier, D., 1988. Frequentist and subjectivist perspectives on the problem of model building in economics (with discussion). Economic Perspectives 2, 121–144.

Poirier, D., 1996. Prior beliefs about fit. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5. Oxford University Press, Oxford, pp. 731–738.

Raftery, A.E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 83, 251–266.

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92, 179–191.

Raftery, A.E., Madigan, D., Volinsky, C.T., 1996. Accounting for model uncertainty in survival analysis improves predictive performance (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5. Oxford University Press, Oxford, pp. 323–349.

Richard, J.F., 1973. Posterior and Predictive Densities for Simultaneous Equation Models. Springer, New York.

Richard, J.F., Steel, M.F.J., 1988. Bayesian analysis of systems of seemingly unrelated regression equations under a recursive extended Natural Conjugate prior density. Journal of Econometrics 38, 7–37.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Smith, M., Kohn, R., 1996. Nonparametric regression using Bayesian variable selection. Journal of Econometrics 75, 317–343.

Smith, A.F.M., Spiegelhalter, D.J., 1980. Bayes factors and choice criteria for linear models. Journal of the Royal Statistical Society, Series B 47, 213–220.

Vandaele, W., 1978. Participation in illegitimate activities; Ehrlich revisited. In: Blumstein, A., Cohen, J., Nagin, D. (Eds.), Deterrence and Incapacitation. National Academy of Sciences Press, Washington, DC, pp. 270–335.

Volinsky, C.T., Madigan, D., Raftery, A.E., Kronmal, R.A., 1997. Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. Applied Statistics 46, 433–448.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P.K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti. North-Holland, Amsterdam, pp. 233–243.

Zellner, A., Siow, A., 1980. Posterior odds ratios for selected regression hypotheses (with discussion). In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics. University Press, Valencia, pp. 585–603.