BENCHMARK PRIORS FOR BAYESIAN MODEL AVERAGING

Carmen Fernández
Dept. of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.

Eduardo Ley
FEDEA, ES-28001 Madrid, Spain

Mark F. J. Steel
Dept. of Economics, University of Edinburgh, Edinburgh EH8 9JY, U.K.

March, 1998
Version: April 2, 1998

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 noninformative 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. In addition, 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 (1995), 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 paper will lead us to propose a "benchmark" prior specification in a linear regression context with model uncertainty.

Keywords. Bayes factors, Markov chain Monte Carlo, Posterior odds, Prior elicitation

JEL Classification System. C11, C15

Address. Mark F. J. Steel, Department of Economics, University of Edinburgh, 50 George Square, Edinburgh EH8 9JY, UK; Tel.: +44-131-650 8352, Fax: +44-131-650 4514, Email: Mark.Steel@ed.ac.uk

The issue of variable selection 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 the topic of model selection has generated in the literature.

Indeed, the issue is an important one, as we are often faced with a situation where a large number of possible regressors can be used and we do not wish to dilute the (often scant) data information we have by including too many regressors. Bayesian methods provide us with a perfectly coherent and interpretable solution, both for selecting and for combining models, through posterior odds. 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.

Broadly speaking, we can distinguish three strands of related literature in this context. Firstly, we mention the fundamentally oriented statistics and econometrics literature on prior elicitation and model selection, 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), Green (1995), Madigan and York (1995), Geweke (1996) and Raftery, Madigan and Hoeting (1997), while Laplace approximations are found in Gelfand and Dey (1995) and Raftery (1996). Finally, there exists a large literature on information criteria, mostly in the context of time series, see *e.g.* Hannan and Quinn (1979), Akaike (1981), Atkinson (1981), Chow (1981). 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 that of Normal linear regression with uncertainty in the choice of regressors. 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 does not violate the rules of probability 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}. 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 model considered. In addition, 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, and will allow researchers in various (applied) fields to use MC^3 techniques on large empirically relevant problems at very modest computing costs. In addition, we present the researcher with a simple diagnostic to assess whether the sampler that generates a Monte Carlo Markov chain over model space has converged.

Section 1 introduces the Bayesian model and the practice of Bayesian model averaging. The prior structure is explained in detail in Section 2, where expressions for Bayes factors are also given. Asymptotic consistency of the latter is studied in Section 3. The setup of the empirical simulation experiment is described in Section 4, while results are provided in the next section. An illustrative example using the economic model of crime from Ehrlich (1973, 1975) concludes the paper.

1. 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 ι_n is an n-dimensional vector of 1's.

C. Fernández gratefully acknowledges financial support from a Training and Mobility of Researchers grant awarded by the European Commission (ERBFMBICT # 961021). C. Fernández and M.F.J. Steel were affiliated to CentER and the Department of Econometrics, Tilburg University, The Netherlands during the early stages of the work on this paper.

This gives rise to 2^k possible sampling models, depending on whether we include or exclude each of the regressors. A model M_j, $j = 1, \ldots, 2^k$, consists of a choice of $0 \leq k_j \leq k$ regressors and leads to

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

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 M_j, namely α, β_j and σ. This distribution will be given through a density function

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

In Section 2, we shall consider specific choices for the density in (1.2) and examine the resulting Bayes factors. In order to complete the prior distribution of the parameters under model M_j, we group the irrelevant components of β under M_j in a vector $\beta_{\sim j} \in \Re^{k-k_j}$. The latter vector follows a Dirac distribution at zero, *i.e.*

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

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 $\mathcal{M}$

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

The Bayesian model is then specified in three consecutive steps:

(1) Through (1.1) we define the distribution of the observables given model M_j and the parameters α, β and σ.

(2) In (1.2)-(1.3) we specify the distribution of the parameters α, β and σ given M_j.

(3) Finally, (1.5) gives the prior probabilities of each of the models.

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

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

This procedure, which is typically referred to as Bayesian model averaging (BMA), is in fact the standard Bayesian procedure under model uncertainty, since it follows directly from the rules of probability calculus upon which the Bayesian paradigm is based [see *e.g.* Leamer (1978), Min and Zellner (1993), Osiewalski and Steel (1993) and Raftery *et al.* (1997)].

Posterior model probabilities are given by

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

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

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

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

Two difficult questions here are how to compute $P(M_j \mid y)$ and how to assess the influence of our prior assumptions on the latter quantity. In cases where $l_y(M_j)$ can be derived analytically, the computation of $P(M_j \mid y)$ is reasonably straightforward applying the MC^3 methodology of Madigan and York (1995). This is a Metropolis algorithm [see *e.g.* Chib and Greenberg (1995)], which allows us to generate drawings from a Markov chain on the model space $\mathcal{M}$ with the posterior model distribution as its stationary distribution. The latter is easily implemented if combined with the choice of a natural conjugate prior structure [see *e.g.* Raftery *et al.* (1997)]. For more complex prior structures that do not allow for an explicit expression for $l_y(M_j)$, the reversible jump methodology of Green (1995) could be applied. An alternative approach was proposed by George and McCulloch (1993, 1997), who do not formally impose zero restrictions as in (1.3), but constrain these coefficients to be concentrated around zero instead (in this way, they get around the problem of a parameter space of varying dimension).

On the other hand, the issue of choosing a "sensible" prior distribution seems further from being resolved. From (1.7) it is clear that the value of $P(M_j \mid y)$ 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 (1997)]. Thus, under little (or under absence of) prior information, the choice of the distribution in (1.2) is a very thorny question. Furthermore, the usual recourse to improper "noninformative" priors does not work in this situation, since the rules of probability no longer apply if we use improper priors on model-specific parameters. Most of the priors that have been proposed in the literature violate the rules of probability calculus, since they are either improper on model-specific parameters [attempts to overcome this are *e.g.* intrinsic Bayes factors as in Berger and Pericchi (1996) or fractional Bayes factors as in O'Hagan (1995)] or are data-dependent through the response variable [as the prior in *e.g.* Raftery *et al.* (1997)]. Here, we will focus on priors that do not have these undesirable properties and are thus, in our view, more suitable for a Bayesian analysis. To this end, we shall propose certain priors and study their behaviour in comparison with other priors previously considered in the literature [*e.g.* in Bernardo (1980) and Laud and Ibrahim (1995, 1996)].

2. 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 next sections, we examine the properties (both finite-sample and asymptotic) of the Bayes factors.

2.1. 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

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

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

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

which corresponds to a Gamma distribution with mean c_0/d_0 and variance c_0/d_0^2 for σ^{-2}. Note that we have assumed a common prior distribution for σ across models. Clearly $m_{0j} \in \Re^{k_j+1}$, V_{0j} a $(k_j + 1) \times (k_j + 1)$ PDS 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

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

where

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

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

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

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

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, App. 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.

2.2. 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

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

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

$$V_{0j}^{-1} = g_{0j} X_j' X_j, \quad (2.8)$$

with $g_{0j} > 0$. From (2.5) it is clear that V_{0j}^{-1} is the prior counterpart of $X_j'X_j$. 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

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

where

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

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 mentioned at the end of Subsection 2.1. In addition, the g-prior in (2.8) can also lead to a prior that is continuously induced across models [see Poirier (1985)] in the sense that the priors for all J 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.

2.3. 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, this takes us outside the rules of probability and 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. It is easy to check that this improper prior leads to 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

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

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

2.4. 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 very much 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 substracting 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):

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

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

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

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

if $k_j \geq 1$ and $k_s \geq 1$. 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 $\overline{y}$, which is eminently reasonable and avoids the sensitivity problems alluded to before. Thus, we shall, henceforth, favour the prior given by the product of (2.11), (2.14) and (2.15).

3. ASYMPTOTIC PROPERTIES AND THE CHOICE OF g_{0j}

Following our comment at the end of Section 2, the remainder of the paper will 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. We note that in (2.16) only g_{0j} remains to be determined. This will be done using a number of properties of the Bayes factor and the posterior model probabilities. In particular, we would like to have consistency (*i.e.* assuming that one of the entertained models is the correct one, we would want the posterior probability of the correct model to converge to one as sample size increases). In addition, we also want sensible behaviour for finite sample sizes, both in terms of posterior model probabilities and predictive ability. In this section we shall focus on large sample results, whereas Section 4 will deal with finite-sample properties through a simulation experiment.

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

$$y = \alpha\iota_n + Z_s\beta_s + \sigma\varepsilon. \tag{3.1}$$

We aim at achieving consistency in the sense that

$$\operatorname*{plim}_{n\to\infty} P(M_s \mid y) = 1 \text{ and } \operatorname*{plim}_{n\to\infty} P(M_j \mid y) = 0 \text{ for all } M_j \neq M_s, \tag{3.2}$$

where the probability limit is taken with respect to the true sampling distribution described in (3.1). 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 (3.2).

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).

The appendix will group some derivations underlying the results in this section. We shall assume throughout that condition $(A.2)$ in the appendix holds. We examine two different functional choices for g_{0j}. Let us first consider dependence on the sample size n and, possibly, on the number of regressors k_j.

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

This is a rather logical choice for g_{0j} in view of the prior in (2.15), which assumes that the prior precision is a fraction g_{0j} of the sample precision. Thus, it seems natural to impose that as sample size increases, the precision of the prior becomes a smaller fraction of that of the sample and vanishes as n goes to infinity. In addition, we let g_{0j} depend on a function w_1 of k_j. The following theorem summarizes our results:

Theorem 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*

$$g_{0j} = \frac{w_1(k_j)}{w_2(n)} \text{ with } \lim_{n\to\infty} w_2(n) = \infty. \tag{3.3}$$

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

$$\lim_{n\to\infty} \frac{w_2'(n)}{w_2(n)} = 0, \tag{3.4}$$

together with either

$$\lim_{n\to\infty} \frac{n}{w_2(n)} \in [0, \infty) \tag{3.5}$$

or

$$w_1(\cdot) \text{ is a nondecreasing function}, \tag{3.6}$$

ensures that the posterior distribution of the models is consistent in the sense defined in (3.2).

On the basis of prior ideas about fit, Poirier (1996) suggests taking $w_2(n) = n$, which satisfies (3.4) and (3.5) and thus leads to consistent Bayes factors. This and other choices will be discussed subsequently.

The proof of Theorem 1 (see appendix) never makes use of the Normality assumption for the error distribution of the 'true' model in (3.1), and, thus, our findings immediately generalize to the case where the components of ε in (3.1) are i.i.d. following any regular distribution 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 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)].

3.2. 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, $D_{js} \geq 1$ regardless of the data [since the numerator in the last factor of (2.16) is then $(y - \overline{y}\iota_n)'(y - \overline{y}\iota_n)$, which is always bigger than or equal to the denominator]. Thus, $P(M_s \mid y)$ 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.* $k_s \geq 1$). See the appendix for proofs.

When M_s is nested within M_j, we have the following result:

$$\operatorname*{plim}_{n\to\infty} D_{js} = 0 \text{ if and only if } w(\cdot) \text{ is an increasing function.} \tag{3.7}$$

The situation becomes less clear-cut when M_s is not nested within M_j. We can show that D_{js} converges to zero when M_j is the model with just the intercept. In addition, taking $w(\cdot)$ to be an increasing function, is sufficient if $k_s \leq k_j$, but we can not assure this if $k_s > k_j \geq 1$ (this will be case-specific). Thus, we can not exclude that models smaller than the true one asymptotically receive positive posterior probability.

On the other hand, if we take $w(k_j)$ to be a constant, we obtain a zero limit for D_{js} if M_s is not nested within M_j. However, in this situation models that nest the true model asymptotically receive positive probability, as follows from (3.7).

3.3. 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 Subsection 3.1 and two consistent information criteria: the Schwarz (or Bayes information) criterion as derived in Schwarz (1978) and the Hannan-Quinn 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:

$$S_{js} = \frac{n}{2} \ln\left(\frac{y' M_{X_s} y}{y' M_{X_j} y}\right) + \frac{k_s - k_j}{2} \ln(n), \tag{3.8}$$

$$HQ_{js} = \frac{n}{2} \ln\left(\frac{y' M_{X_s} y}{y' M_{X_j} y}\right) + \frac{k_s - k_j}{2} C_{HQ} \ln\ln(n). \tag{3.9}$$

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 1, can be characterized by the following result:

Theorem 2. *Consider the Bayesian model described in Theorem 1, with g_{0j} verifying* (3.4) *together with either* (3.5) *or* (3.6)*. Then the Bayes factor in* (2.16) *satisfies:*

$$\operatorname{plim} \frac{\ln B_{js}}{\frac{n}{2}\ln\left(\frac{y'M_{X_s}y}{y'M_{X_j}y}\right)+\frac{k_s-k_j}{2}\ln w_2(n)} = 1, \tag{3.10}$$

where the probability limit is taken with respect to the model M_s as described in (3.1).

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

$$\operatorname{plim} \frac{\ln B_{js}}{S_{js}} = 1, \tag{3.11}$$

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

$$\operatorname{plim} \frac{\ln B_{js}}{HQ_{js}} = 1. \tag{3.12}$$

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 1 does not verify (3.5), which is why we impose that $w_1(\cdot)$ fulfills (3.6). 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.

As a final note, it is again worth mentioning that Theorem 2 also holds if the error terms in (3.1) follow a non-Normal distribution.

4. THE SIMULATION EXPERIMENT

4.1. Introduction

In this section we perform a simulation experiment to assess the performance of different choices of g_{0j} in finite sampling. In addition to Bayes factors, we will compute posterior model probabilities and evaluate predictive ability under several choices of g_{0j}. Our results in this section 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

$$P(M_j) = p_j = 2^{-k}, \quad j = 1, \ldots, k. \tag{4.1}$$

Creating the design matrix of the simulation experiment 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

$$(r_{(11)}, \ldots, r_{(15)}) = (r_{(1)}, \ldots, r_{(5)})(.3 \quad .5 \quad .7 \quad .9 \quad 1.1)'(1 \quad 1 \quad 1 \quad 1 \quad 1) + E \tag{4.2}$$

where E is an $n \times 5$ matrix of independent standard Normal deviates. Note that (4.2) 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). Curiously, this correlation structure differs from the one reported in Raftery *et al.* (1997), which seems in conflict with (4.2). 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

$$\text{Model 1}: \quad y = 4 + 2z_{(1)} - z_{(5)} + 1.5z_{(7)} + z_{(11)} + 0.5z_{(13)} + u, \tag{4.3}$$

$$\text{Model 2}: \quad y = 1 + u, \tag{4.4}$$

where the elements of u are independently Normally distributed with mean zero and variance $\sigma^2 = 6.25$. Whereas Model 1 is meant to capture a more or less realistic situation where one third of the regressors intervene, 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.

4.2. Choices for g_{0j}

Based on the theory in Section 3, we shall consider seven different choices for g_{0j}. From Theorem 1, priors a-f all lead to consistency, in the sense of asymptotically selecting the correct model. In addition, from Corollary 1, the log of Bayes factors obtained under priors a-c behave asymptotically like the Schwarz criterion, whereas those obtained under priors e and f respectively behave like the Hannan-Quinn criterion with $C_{HQ} = 3$ and $C_{HQ} = 1$. Prior d provides an intermediate case in terms of asymptotic penalty for large models.

Prior a: $g_{0j} = \frac{1}{n}$

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 Wasserman (1995) and the g-prior (using a Cauchy prior on β given σ) used in Zellner and Siow (1980). In addition, this choice of g_{0j} does not depend on j and, thus, implies continuously induced priors across models, in the sense of Poirier (1985).

Prior b: $g_{0j} = \frac{k_j}{n}$

Here we assign more information to the prior as we have more regressors in the model, *i.e.* we expect the sample information to be more diluted as the number of regressors grows.

Prior c: $g_{0j} = \frac{k^{1/k_j}}{n}$

Now prior information decreases with the number of regressors in the model.

Prior d: $g_{0j} = \sqrt{\frac{k_j}{n}}$

This is an intermediate case, where we choose $w_2(n) = n^{1/2}$, and we have a smaller asymptotic penalty term for large models than in the Schwarz criterion [see (3.8)].

Prior e: $g_{0j} = \frac{1}{(\ln n)^3}$

Here we choose $w_2(n)$ so as to mimic the Hannan-Quinn criterion in (3.9) with $C_{HQ} = 3$ as n becomes large. This prior is also continuously induced across models.

Prior f: $g_{0j} = \frac{\ln(k_j+1)}{\ln n}$

Now $w_2(n)$ increases even slower with sample size and we have asymptotic convergence of $\ln B_{js}$ to HQ_{js} for $C_{HQ} = 1$.

Prior g: $g_{0j} = \frac{\delta\gamma^{1/k_j}}{1-\delta\gamma^{1/k_j}}$

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 γ 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 $k = 15$ this implies: $\gamma = 0.64889, \delta = 0.15411$. Note that this prior choice does not lead to consistency if the data are generated from a model with only the intercept, the "null model" in (4.4) (see Subsection 3.2). If $\gamma < 1$ then g_{0j} is an increasing function of k_j, and following Subsection 3.2 we know that $\text{plim}_{n\to\infty} B_{js} = 0$ when $k_s \geq 1$ and M_j is either the null model or $k_j \geq k_s$. In cases when $1 \leq k_j < k_s$ it depends on $(A.9)$ whether B_{js} converges to zero or not.

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 [see 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 of which we substract 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 explicit the conditioning on the regressors in z_f and Z in the notation. In accordance with the expression in (1.6), the out-of-sample predictive distribution for $f = 1, \ldots, q$ will be characterized by

$$\begin{aligned} p(y_f \mid z_f, y, Z) = &\sum_{j=1}^{J} f_S^1(y_f \mid n-1, \overline{y} + \frac{1}{g_{0j}+1} z'_{f,j}\beta_j^*, \\ &\frac{n-1}{d_j^*}\{1 + \frac{1}{n} + \frac{1}{g_{0j}+1} z'_{f,j}(Z_j'Z_j)^{-1}z_{f,j}\}^{-1})P(M_j \mid y, Z), \end{aligned} \tag{4.5}$$

where $\overline{y}$ 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

$$d_j^* = \frac{1}{g_{0j}+1} y' M_{X_j} y + \frac{g_{0j}}{g_{0j}+1}(y - \overline{y}\iota_n)'(y - \overline{y}\iota_n) \tag{4.6}$$

The term in (4.5) 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). This is the first predictive criterion we will compute. For each value of z_f we shall generate a number, say v, of responses from the underlying true model [(4.3) or (4.4)] and base our predictive measure on (4.5) evaluated in these out-of-sample observations $y_{f1}, \ldots, y_{fv}$, namely:

$$LPS(z_f, y, Z) = -\frac{1}{v}\sum_{i=1}^{v} \ln p(y_{fi} \mid z_f, y, Z), \tag{4.7}$$

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, Gavrin and Raftery (1995) give an interpretation for differences in log predictive scores in terms of one toss with a biased coin.

More formally, the criterion in (4.7) 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 $p(y_f \mid z_f)$ in (4.3) or (4.4) and the out-of-sample predictive density in (4.5) can be written as

$$\begin{aligned} KL\{p(y_f \mid z_f), p(y_f \mid z_f, y, Z)\} = &\int_{\Re} \{\ln p(y_f \mid z_f)\} p(y_f \mid z_f) dy_f - \\ &\int_{\Re} \{\ln p(y_f \mid z_f, y, Z)\} p(y_f \mid z_f) dy_f, \end{aligned} \tag{4.8}$$

where the first integral is the negative entropy of the sampling density, and the second integral can be seen as a theoretical counterpart of (4.7) 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_{fi} 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 (4.3), regardless of z_f. By the nonnegativity of the Kullback-Leibler divergence, this constitutes a lower bound for $LPS(z_f, y, Z)$ of 2.335.

We now have a measure of predictive performance, $LPS(z_f, y, Z)$, for each $z_f, f = 1, \ldots, q$. We can, of course, immediately assess the distribution of this quantity as z_f varies, through its empirical distribution corresponding to all q generated values of z_f.

Finally, we can investigate the calibration of the predictive and compare the entire predictive density function in (4.5) with the known sampling distribution of the response in (4.3) or (4.4) 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 can not 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. 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 (4.3), the one leading to the median value and the one giving rise to the largest value. For the sampling model in (4.4), the value of z_f does not intervene, so here we shall just present a graph of the predictive for one (randomly chosen) value z_f (the sampling model now does not change with z_f, but the predictive does, as long as we assign nonzero probability to models larger than the null model). In addition, we shall present properties of LPS and the predictive coverage averaged 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. FINITE SAMPLE 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 $\mathcal{M}$ consisting of the current state of the chain, say M_s, and all models containing one regressor more or 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 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. Lee (1996) introduces this idea 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 don't require any ergodic properties. 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 almost the same answers with empirical model frequencies. This resulted in Markov Chain Monte Carlo with 50,000 recorded drawings after a burn-in of 20,000 drawings. A useful diagnostic to assess convergence of the Markov chain is the correlation coefficient of the exact Bayes factors computed through (2.16) and the relative frequencies of model visits.

In order to avoid results depending on the particular sample analyzed, 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 used in the simulation will be $n = 50$, 100, 1000, 10,000 and 100,000. Furthermore, we generate $q = 19$ different vectors of regressors z_f for the forecasts of Model 1, 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.

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 evaluated models in order to reduce the number of calculations. The entire simulation was run on one single 120MHz 604 PowerPC-based desktop computer. On a PowerMacintosh 7600, each 20,000–50,000 chain would take an average (over priors) time in seconds of: 209, 58, 5, 18, and 117; for $n = 50$, 100, 1000, 10,000 and 100,000. The source code is posted at http://econwpa.wustl.edu and it is freely available.

5.2 Posterior model inference

Section 3 contains some results concerning the asymptotic behaviour of our Bayesian analysis as the number of observations n goes to infinity. Here we summarize the main results for various finite values of n.

5.2.1. Results under Model 1

One of the main 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 five sample sizes used and rows order the different priors a through g introduced in Subsection 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 $3.052 \cdot 10^{-5}$. We know from the theoretical results in Subsection 3.1 that priors a-f are consistent in the sense of (3.2). From Subsection 3.2, we remain inconclusive about consistency under prior g, since we cannot exclude that models with less regressors than the true one receive, asymptotically, positive posterior probability. However, our simulation results will suggest that consistency holds in our particular example. It is clear from Table 1 that the posterior probability of Model 1 varies greatly in finite samples. Whereas prior d already performs very well for $n = 1000$, getting average probabilities of the correct model upwards of 0.97, prior e only obtains a probability of 0.60 with a sample as large as 100,000. 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 f 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 1. Model 1: Means and Stds of the posterior probability of the true model.

n	50		100		1000		10,000		100,000	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	0.0128	0.0197	0.0575	0.0618	0.5293	0.1401	0.8111	0.0928	0.9254	0.0760
b	0.0066	0.0091	0.0332	0.0338	0.4407	0.1373	0.7601	0.1064	0.9048	0.0841
c	0.0110	0.0159	0.0519	0.0533	0.4860	0.1374	0.7853	0.0999	0.9145	0.0804
d	0.0029	0.0026	0.0205	0.0188	0.9730	0.0196	1.0000	0.0000	1.0000	0.0000
e	0.0141	0.0223	0.0586	0.0616	0.3610	0.1251	0.5139	0.1327	0.5981	0.1327
f	0.0020	0.0014	0.0128	0.0107	0.7762	0.3421	1.0000	0.0000	1.0000	0.0000
g	0.0026	0.0026	0.0069	0.0056	0.2773	0.0864	1.0000	0.0000	1.0000	0.0000

Table 2. Model 1: Quartiles of ratio of posterior probabilities; True Model vs 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	1.5	3.2	6.3	3.0	5.8	8.6	9.1	19.0	29.8	27.6	67.8	90.5	79.0	193.4	285.5
b	1.1	2.6	4.2	2.2	4.1	6.2	9.6	16.6	22.1	16.6	36.3	66.3	59.7	113.4	194.7
c	1.2	3.4	5.8	2.0	5.1	8.3	7.0	13.8	22.9	26.3	55.9	73.6	62.3	135.6	236.1
d	1.6	2.7	3.5	1.9	3.8	5.8	226.5	416.4	629.4	∞	∞	∞	∞	∞	∞
e	1.7	4.0	7.0	2.0	4.4	8.7	4.7	9.2	16.5	9.7	19.7	24.9	12.7	22.2	34.2
f	1.4	2.3	3.3	1.4	3.9	5.1	11.4	238.7	3625.8	∞	∞	∞	∞	∞	∞
g	1.2	2.3	2.8	1.9	2.8	3.9	5.9	10.6	12.9	∞	∞	∞	∞	∞	∞

Table 3. Model 1: Means and Stds of Number of Models Visited.

n	50		100		1000		10,000		100,000	
Prior	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
d	7159	1549	2810	838	15	4	1	0	1	0
e	2056	596	1158	301	237	47	151	37	110	24
f	8677	1608	3555	1204	3	1	1	0	1	0
g	5480	1353	3322	809	654	89	1	0	1	0

Table 3 records means and standard deviations of the number of visited models in the 50,000 recorded drawings of the chain in model space. 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 f, and never less than six percent on average (Prior a). The sampler visits less models as n increases, and for $n = 1000$ we already have very few visited

models for priors f in particular and also for d. When 10,000 observations are available, that is enough to make the sampler stick to one model (the correct one) for priors d, f and g. Surprisingly, whereas prior g 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. Finally, note that even with 100,000 observations, prior e still makes the sampler visit almost 110 models on average.

Table 4. Model 1: Means and Stds of Posterior Probabilities of Including each regressor.

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

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 (4.9), Model 1 contains regressors 1,5,7,11 and 13 (indicated with arrows in Table 4). To save space, we shall only report this for $n = 50$ and $n = 1000$. 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 coefficients of opposite signs. The posterior probabilities of including regressors not contained in the correct model is relatively small. What is not clearly exemplified by Table 4 is that priors a through f tend to choose alternatives that are nested by Model 1 for small sample sizes, whereas prior g puts 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 f 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 d and the largest for prior g). Alternative models tend to nest the correct model for all priors, except prior f, 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 (4.4), 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 and most priors lead to small probabilities of selecting the correct model.

Overall, priors a and e do best for small sample sizes, whereas larger sample sizes are most favourable to priors a and b. Despite these rather small 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 the ratio of the posterior probabilities of Model 2 and the best other model are presented. Only prior f leads to a first quartile below unity. Overall, priors a and b seem to do best on this criterion. The difficulty of pinning down the correct (null) model can also be inferred from Table 7, where means and standard deviations of the number of visited models are presented. It is clear that some priors (like b, d, f and g) make the chain wander a lot for small sample sizes. Priors f and g retain this problematic behaviour even for sample sizes as large as 100,000. Interestingly, whereas prior f leads to (very slow) improvements as n increases, the bad behaviour with prior g seems entirely unaffected by sample size. Of course, we know from the theory in Subsection 3.2 that prior g does not lead to consistent Bayes factors in this case. The number of models visited is relatively small for prior a, which seems to emerge as the winner from the posterior results under Model 2.

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

n	50		100		1000		10,000		100,000	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	0.0320	0.0269	0.0722	0.0494	0.3812	0.1543	0.7199	0.1346	0.8995	0.0529
b	0.0021	0.0028	0.0114	0.0124	0.2606	0.1506	0.6910	0.1441	0.8960	0.0568
c	0.0099	0.0081	0.0238	0.0157	0.1494	0.0715	0.4148	0.1201	0.7080	0.1009
d	0.0003	0.0003	0.0006	0.0006	0.0066	0.0066	0.0427	0.0300	0.1570	0.0938
e	0.0407	0.0322	0.0764	0.0492	0.2216	0.1050	0.3569	0.1220	0.4733	0.1235
f	0.0001	0.0002	0.0002	0.0002	0.0005	0.0006	0.0009	0.0008	0.0014	0.0015
g	0.0010	0.0012	0.0011	0.0012	0.0014	0.0014	0.0013	0.0011	0.0014	0.0014

Table 6. Model 2: Quartiles of ratio of posterior probabilities; True Model vs 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
d	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
e	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
f	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
g	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

Table 7. Model 2: Means and Stds of Number of Models Visited.

n	50		100		1000		10,000		100,000	
Prior	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
a	3921	937	2213	737	445	149	116	27	37	10
b	14482	2379	11651	1589	1813	817	245	90	54	17
c	4183	781	2496	523	513	163	159	28	60	12
d	16869	2221	15784	2097	10251	2022	4852	1576	1960	690
e	3529	729	2396	539	948	279	513	127	329	79
f	17994	2105	17601	2133	16381	1879	14963	3065	13364	3671
g	12587	1561	12580	1483	12653	1353	12532	2201	12413	2205

In summary, the posterior results for Model 1 point towards prior d as the best choice for most practical purposes, whereas prior a seems preferable for small samples and when the null model has generated the data.

5.3. Predictive inference

5.3.1. Results under Model 1

As discussed in Subsection 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}.

Table 8. 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	2.326	2.325	2.338	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	2.326	2.325	2.338	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	2.326	2.325	2.338	2.331	2.350	2.345
d	2.691	2.448	2.475	2.507	2.406	2.410	2.358	2.356	2.354	2.333	2.326	2.338	2.332	2.351	2.345
e	2.474	2.428	2.428	2.393	2.389	2.391	2.333	2.355	2.347	2.326	2.325	2.338	2.331	2.350	2.345
f	2.836	2.470	2.530	2.636	2.423	2.463	2.475	2.378	2.412	2.440	2.355	2.385	2.417	2.362	2.379
g	2.492	2.430	2.440	2.450	2.392	2.395	2.418	2.362	2.381	2.419	2.349	2.374	2.422	2.364	2.382

Firstly, Table 8 presents the median of $LPS(z_f, y, Z)$ as computed in (4.7) 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 in Subsection 4.3. Of course, LPS in (4.7) is only a Monte Carlo approximation to this integral (based on a mere 100 drawings), so this lower bound is not always strictly adhered to. Under priors a through e we are predicting the sampling density virtually exactly with samples of size $n = 1000$ or more. Of these five priors, prior d performs slightly worse for very small samples (in particular, for the z_f corresponding to the minimum sampling mean). Priors f and g tend to be further from the actual sampling density and do not lead to perfect prediction even with 100,000 observations. Prior f, in particular, leads to rather large values for LPS when n is 1000 or smaller and conditionally upon z_{min}.

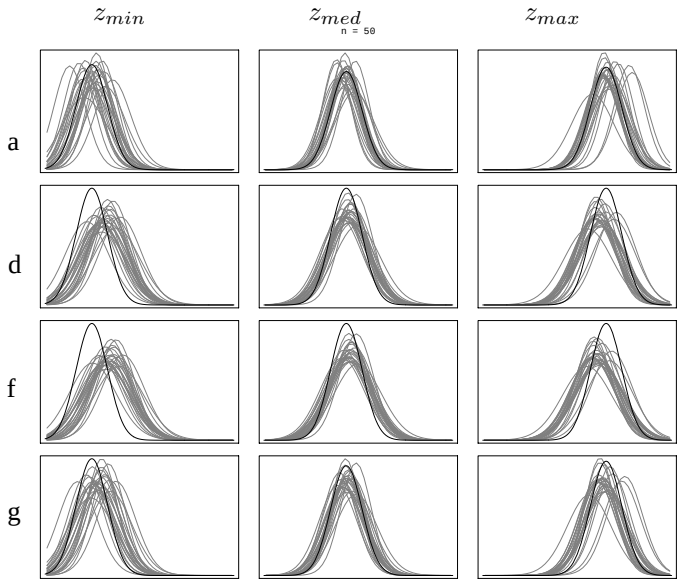

Fig. 1. Model 1: Predictive densities, $n = 50$.

In order to find out more about the differences between the predictive density in (4.5) and the sampling density in (4.3), we can overplot both densities for the three values of z_{min}, z_{med} and z_{max}. Figures 1 and 2 display this comparison for different values of n and the predictives for 25 of the 100 generated samples (to avoid cluttering the graphs). The dark line corresponds to the actual sampling density. Since the predictives from priors a-c and e are very close for all sample sizes, we shall only present the graphs for priors a, d, f and g. 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, b, c and e 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 (in between 0.02 and 0.08). The priors d, f and g 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,

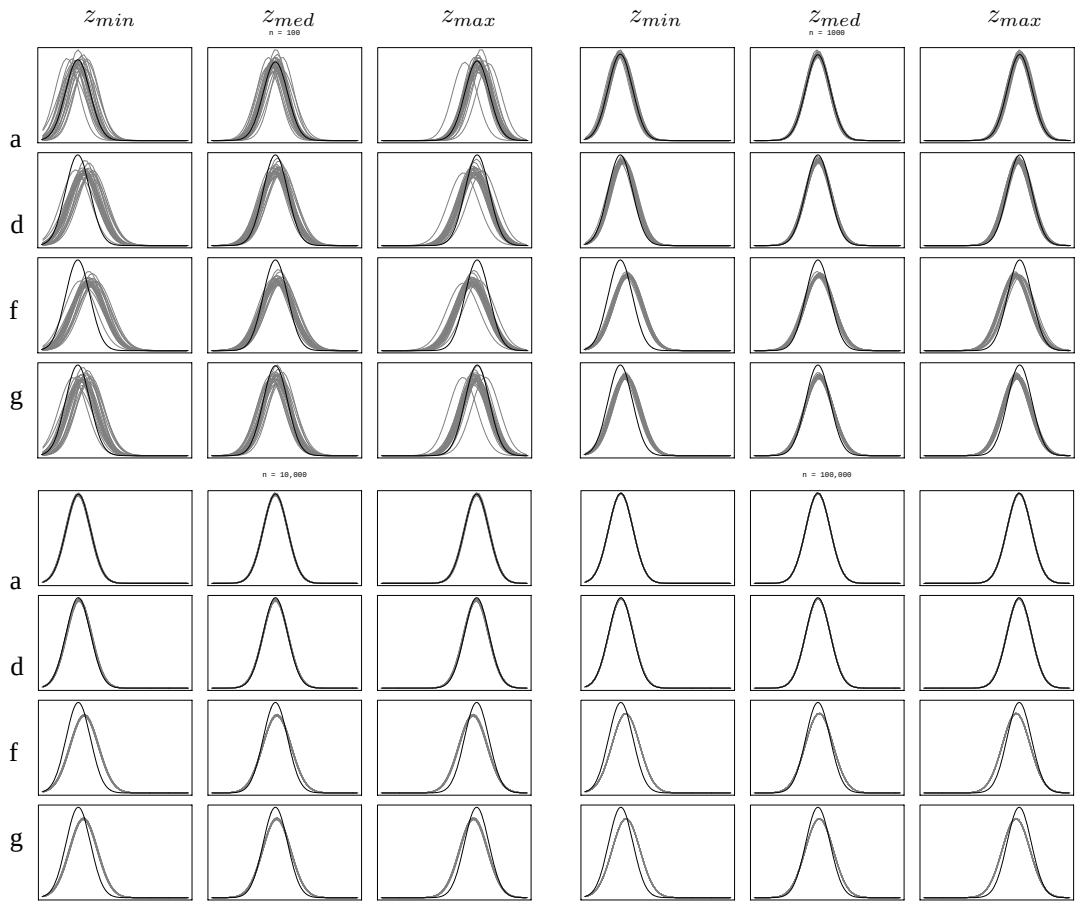

Fig. 2. Model 1: Predictive densities, $n = 100$, 1000, 10,000, and 100,000.

less concentrated than the sampling density. This behaviour can easily be understood once we realize that the locations of each of the components in (4.5) (the posterior mean of $\alpha + z_f'\beta$ under each of the models considered) are clearly shrunk more towards the sample mean $\bar{y}$ 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 d, f and g. 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 d (g_{0j} is then equal to 0.06) whereas it persists for priors f and g even with 100,000 observations (where g_{0j} takes the values 0.14 and 0.16, respectively).

Table 9. 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
d	2.473	2.416	2.347	2.335	2.334
e	2.428	2.382	2.339	2.334	2.333
f	2.502	2.452	2.393	2.375	2.363
g	2.433	2.452	2.369	2.366	2.366

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 9, where the medians (computed across the 1900 sample-z_f combinations) are recorded for the different priors and sample sizes. Clearly, whereas all priors except for prior f (and, to a lesser extent, prior d) lead to comparable predictive behaviour for very small n, the fact that g_{0j} is constant in n makes prior g lose ground with respect to the other priors as n increases. Prior f always performs worse than priors a through e. Note that

priors a through e lead to median LPS values that are roughly equal to the theoretical minimum of 2.335, implying perfectly accurate prediction, for $n \geq 1000$.

Alternatively, we can compare the percentiles of the sampling distribution and the predictive in (4.5). We compute the predictive percentiles corresponding to the $1^{st}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}$, and 99^{th} sampling percentile. The quartiles of these numbers, calculated over all 1900 sample-z_f combinations, are presented in Table 10. This confirms that priors a-c and e lead to better predictions for small sample sizes, where the predictives from d,f and g are too spread out. Starting at $n = 1000$, prior d predicts well, whereas the inaccurate predictions with priors f and g persist even for very large sample sizes. Comparing Figure 2 with Table 10, it is clear that most of the spread in the percentiles for priors f and g with $n \geq 10,000$ is due to the bias toward the median (shrinkage). Remember that Table 10 averages over the 19 different values of z_f.

5.3.2. Results under Model 2

As mentioned in Subsection 5.2.2, it is very 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 Table 10, where predictive percentiles are compared with the actual sampling percentiles. Clearly, the incorrectly chosen models are such that they do not lead our predictions (averaged over all the chosen models as in (4.5)) far astray. Table 10 contains predictive percentiles for $n = 50$ and 1000, and 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.

Summarizing the predictive performance, we can state the following. For Model 1 priors a, b, c, and e (which induce the smallest values for g_{0j}) seem to do best, but for 1000 or more observations, prior d does just as well. All these priors predict virtually exactly for $n = 1000$ or more. Priors f and g imply too much shrinkage, as a result of large values for g_{0j} (in the case of prior g, this value does not depend on n at all), and thus do worse in prediction. In the case of Model 2 (the null model), prediction is virtually perfect under all priors, even with small samples. For this model the issue of shrinkage is, of course, less problematic.

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 (1991), 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 35-39, 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 a 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 (0.9896) between 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 (BARS, as explained in Subsection 5.1). In all, 3378 different models were visited, and the best 10% of those models account for 73.5% of the posterior model probability. Thus, posterior mass is not highly concentrated on just a few models. Note that this run takes a mere 80 seconds on a 120MHz 604 PowerPC Macintosh personal computer.

Table 11 presents the 9 models that receive over 1% posterior probability. The best model is the same as that in Raftery *et al.* (1997). In general, model probabilities are very similar, even though our prior is quite

Table 10. Quartiles of the Predictive Percentiles.

	Model 1: Prior a															Model 2: Prior a					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.005	.012	.026	.005	.010	.018	.008	.010	.012	.009	.010	.011	.010	.010	.010	.005	.010	.019	.009	.010	.011
.05	.024	.052	.094	.029	.049	.074	.045	.050	.058	.048	.050	.052	.049	.050	.050	.031	.048	.073	.046	.050	.054
.25	.143	.237	.346	.174	.237	.305	.234	.251	.271	.245	.250	.255	.248	.249	.251	.194	.241	.299	.239	.249	.259
.50	.339	.471	.600	.391	.479	.565	.480	.502	.526	.494	.500	.507	.498	.500	.502	.432	.496	.559	.489	.498	.510
.75	.587	.715	.816	.650	.729	.798	.734	.752	.771	.745	.751	.755	.748	.750	.752	.691	.751	.804	.741	.750	.758
.95	.871	.932	.965	.906	.941	.963	.944	.951	.957	.948	.950	.952	.950	.950	.951	.922	.949	.969	.947	.950	.954
.99	.959	.982	.992	.975	.986	.993	.988	.990	.992	.990	.990	.991	.990	.990	.990	.979	.989	.995	.989	.990	.991
	Model 1: Prior b															Model 2: Prior b					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.008	.018	.036	.007	.013	.022	.009	.010	.012	.009	.010	.011	.010	.010	.010	.004	.010	.021	.009	.010	.011
.05	.032	.064	.107	.035	.054	.079	.045	.051	.058	.048	.050	.052	.049	.050	.050	.027	.047	.079	.045	.050	.054
.25	.154	.242	.339	.179	.240	.306	.233	.251	.271	.245	.250	.255	.248	.249	.251	.175	.236	.320	.237	.248	.260
.50	.333	.453	.569	.385	.469	.549	.479	.501	.524	.494	.500	.507	.498	.500	.502	.405	.491	.586	.487	.498	.512
.75	.562	.679	.775	.630	.709	.775	.731	.750	.769	.745	.751	.755	.748	.750	.752	.669	.749	.819	.740	.750	.760
.95	.836	.903	.944	.891	.927	.953	.943	.950	.956	.948	.950	.952	.950	.950	.951	.916	.950	.972	.946	.950	.954
.99	.940	.968	.985	.967	.981	.989	.988	.990	.991	.989	.990	.991	.990	.990	.990	.978	.989	.995	.989	.990	.991
	Model 1: Prior c															Model 2: Prior c					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.005	.013	.028	.006	.011	.019	.009	.010	.012	.009	.010	.011	.010	.010	.010	.006	.011	.019	.009	.010	.011
.05	.025	.053	.096	.030	.050	.075	.045	.051	.058	.048	.050	.052	.049	.050	.050	.032	.049	.074	.046	.050	.054
.25	.145	.238	.346	.176	.238	.307	.234	.252	.271	.245	.250	.255	.248	.249	.251	.195	.242	.300	.238	.249	.260
.50	.339	.470	.601	.391	.479	.565	.479	.502	.525	.494	.500	.507	.498	.500	.502	.432	.496	.559	.488	.498	.510
.75	.586	.712	.812	.649	.729	.797	.733	.752	.771	.745	.751	.755	.748	.750	.752	.690	.751	.803	.740	.750	.759
.95	.868	.929	.963	.905	.939	.962	.944	.951	.957	.948	.950	.952	.950	.950	.951	.922	.949	.968	.946	.950	.954
.99	.957	.981	.992	.974	.986	.992	.988	.990	.992	.990	.990	.991	.990	.990	.990	.979	.989	.995	.989	.990	.991
	Model 1: Prior d															Model 2: Prior d					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.012	.026	.050	.011	.019	.033	.011	.013	.017	.010	.011	.012	.010	.010	.011	.005	.011	.020	.009	.010	.012
.05	.040	.075	.126	.041	.064	.099	.049	.058	.067	.050	.052	.055	.050	.051	.051	.029	.049	.077	.044	.050	.055
.25	.156	.236	.337	.173	.233	.313	.230	.253	.278	.243	.250	.259	.247	.250	.252	.185	.241	.312	.234	.248	.263
.50	.310	.420	.539	.351	.436	.530	.458	.486	.517	.485	.495	.505	.495	.498	.501	.419	.494	.573	.480	.498	.516
.75	.508	.624	.730	.571	.657	.738	.700	.725	.750	.733	.741	.749	.744	.747	.750	.680	.750	.811	.735	.749	.763
.95	.777	.855	.911	.836	.885	.926	.924	.934	.944	.942	.945	.948	.948	.949	.950	.920	.949	.970	.945	.950	.955
.99	.896	.941	.969	.937	.961	.978	.981	.984	.987	.987	.988	.989	.989	.990	.990	.979	.989	.995	.989	.990	.991
	Model 1: Prior e															Model 2: Prior e					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.005	.012	.026	.006	.010	.018	.009	.010	.012	.010	.010	.011	.010	.010	.010	.006	.011	.019	.009	.010	.011
.05	.023	.051	.094	.029	.049	.075	.045	.051	.058	.048	.050	.052	.049	.050	.050	.032	.049	.074	.046	.050	.054
.25	.143	.236	.346	.174	.237	.306	.233	.251	.271	.245	.250	.256	.248	.249	.251	.196	.241	.299	.238	.248	.260
.50	.338	.471	.600	.392	.480	.566	.479	.502	.526	.494	.500	.506	.497	.500	.502	.434	.496	.558	.487	.498	.511
.75	.587	.715	.817	.650	.730	.798	.732	.751	.770	.744	.750	.755	.748	.750	.752	.691	.751	.803	.740	.750	.759
.95	.871	.932	.966	.906	.941	.963	.943	.950	.957	.948	.950	.952	.950	.950	.951	.921	.948	.968	.946	.950	.954
.99	.960	.982	.992	.975	.986	.993	.988	.990	.992	.989	.990	.991	.990	.990	.990	.979	.989	.995	.989	.990	.991
	Model 1: Prior f															Model 2: Prior f					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.013	.029	.057	.012	.022	.041	.013	.020	.031	.012	.017	.026	.012	.016	.023	.006	.011	.020	.009	.010	.012
.05	.042	.078	.136	.041	.068	.113	.049	.068	.097	.049	.064	.087	.048	.062	.081	.031	.049	.075	.045	.050	.055
.25	.150	.232	.343	.163	.231	.324	.199	.246	.310	.204	.245	.300	.209	.246	.293	.191	.241	.308	.234	.248	.263
.50	.296	.403	.532	.325	.415	.528	.389	.451	.526	.403	.456	.523	.415	.464	.520	.425	.495	.567	.481	.498	.516
.75	.483	.599	.717	.528	.623	.724	.610	.669	.734	.631	.681	.738	.648	.693	.740	.684	.750	.805	.735	.749	.763
.95	.742	.831	.898	.794	.857	.912	.861	.893	.924	.879	.905	.930	.892	.913	.934	.922	.949	.968	.945	.950	.955
.99	.870	.925	.960	.909	.944	.970	.950	.965	.977	.961	.971	.981	.967	.975	.982	.979	.989	.994	.989	.990	.991
	Model 1: Prior g															Model 2: Prior g					
	$n = 50$			$n = 100$			$n = 1000$			$n = 10,000$			$n = 100,000$			$n = 50$			$n = 1000$		
%	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
.01	.009	.020	.039	.010	.017	.029	.013	.017	.024	.012	.016	.023	.012	.016	.023	.005	.010	.020	.009	.010	.012
.05	.034	.067	.112	.040	.061	.093	.051	.064	.082	.049	.062	.082	.048	.062	.082	.028	.048	.076	.045	.050	.055
.25	.157	.244	.341	.180	.239	.312	.214	.250	.292	.210	.246	.294	.208	.246	.295	.182	.238	.312	.234	.248	.263
.50	.332	.451	.566	.372	.452	.536	.420	.465	.518	.415	.461	.520	.411	.462	.520	.416	.494	.575	.481	.499	.516
.75	.554	.672	.768	.598	.680	.751	.651	.693	.737	.646	.689	.739	.643	.690	.740	.678	.750	.812	.735	.749	.763
.95	.827	.896	.938	.860	.904	.937	.892	.912	.931	.890	.911	.932	.888	.911	.933	.919	.950	.971	.945	.950	.955
.99	.934	.964	.982	.951	.971	.983	.966	.974	.981	.966	.974	.982	.965	.974	.982	.978	.989	.995	.989	.990	.991

Table 11. Models with more than 1% posterior probability.

	Prob.	Included Regressors									
1	2.55%	1	3	4		9	11		13	14	
2	2.48%	1	3	4		9	11		13	14	15
3	1.68%	1	3		5	9	11		13	14	
4	1.52%	1	3	4			11		13	14	
5	1.41%	1	3	4	8	9	11		13	14	
6	1.28%	1	3	4		9			13	14	15
7	1.11%	1	3	4		9	11	12	13	14	15
8	1.04%	1	3		5	9	11		13	14	15
9	1.02%	1	3		5		11		13	14	

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 results in Section 5 leads to results that are very close to those with the rather carefully and laboriously elicited prior of Raftery *et al.* (1997). In addition, the latter prior depends on the data, as mentioned in Subsection 2.3.

Table 12. Posterior Probabilities of Including each Regressor.

	Regressor	Prob.
1	Percentage of males age 14–24	85.95 %
2	Indicator variable for southern state	22.46 %
3	Mean years of schooling	98.77 %
4	Police expenditure in 1960	66.78 %
5	Police expenditure in 1959	41.58 %
6	Labor force participation rate	14.74 %
7	Number of males per 1,000 females	15.24 %
8	State population	32.67 %
9	Number of nonwhites per 1,000 people	68.59 %
10	Unemployment rate for urban males, age 14–24	20.29 %
11	Unemployment rate for urban males, age 25–39	60.63 %
12	Wealth	30.79 %
13	Income inequality	99.94 %
14	Probability of imprisonment	90.73 %
15	Average time served in prisons	33.10 %

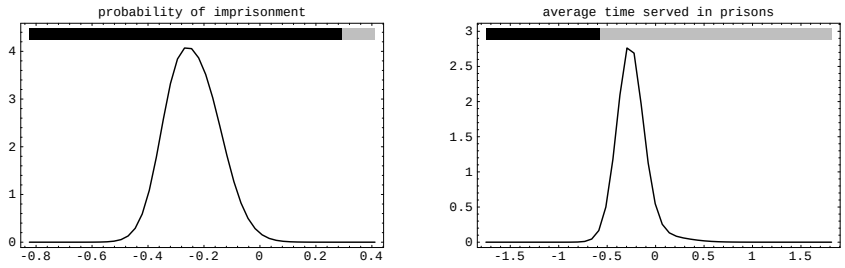

Fig. 3. Posterior density functions: Regressors 14 and 15.

Posterior probabilities of including each of the regressors are given in Table 12, which clearly indicates that schooling and 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 12 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 posterior density functions (averaging over models with posterior probabilities) are given in Figure 3. The coefficients of these regressors can be interpreted as elasticities. The gauge on top indicates (in black) the posterior probability of inclusion. 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

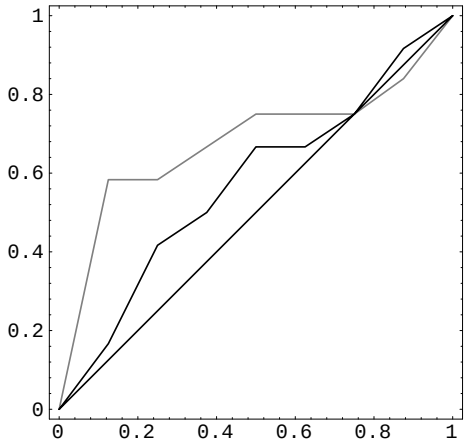

Fig. 4. Q-Q plot with 75%–25% sample split.

inclusion of one third (see also Table 13).

If we split the data randomly into 35 observations used for estimation and 12 to be predicted, we obtain the predictive Q-Q plot in Figure 4. This is what Raftery *et al.* (1997) call a "calibration plot". We note that the best single model (graph in grey) predicts considerably worse than the predictive in (4.5) resulting from BMA (graph in black). Averaging over models with posterior probabilities does a much better job at predicting the 12 remaining observations than simply taking the model with highest posterior probability.

7. RECOMMENDATIONS

The prior structure we have proposed in Section 2 only requires the choice of one scalar hyperparameter, called g_{0j}. We make g_{0j} a possible function of the sample size and of the number of regressors in the model under consideration, M_j. 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 seem to indicate that the following two priors are reasonable choices:

- prior a, where $g_{0j} = 1/n$, for small n and data generated from models with relatively few regressors
- prior d, where $g_{0j} = \sqrt{k_j/n}$, in other cases.

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

Our empirical simulation results compare favourably to those reported in George and McCulloch (1993) and Raftery *et al.* (1997), whereas our prior does not depend on the response variable and is very easy to elicit.

APPENDIX: SOME ASYMPTOTIC RESULTS

Before examining consistency, we need to establish some preliminary results. Some of these results are not new, whereas others are easy to derive. Thus, we do not present the proof of Lemma 1.

Lemma 1. *Under the sampling model M_s in* (3.1)*,*
(i) If M_s is nested within or is equal to model M_j,

$$\operatorname*{plim}_{n\to\infty} \frac{y' M_{X_j} y}{n} = \sigma^2. \tag{A.1}$$

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

$$lim_{n\to\infty} \frac{(\alpha, \beta_s')X_s' M_{X_j} X_s (\alpha, \beta_s')'}{n} = b_j \in (0, \infty), \tag{A.2}$$

we obtain

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

A.1. Proof of Theorem 1

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

$$C_{js} = \left(\frac{w_1(k_j)}{g_{0j}+1}\right)^{k_j/2} \left(\frac{g_{0s}+1}{w_1(k_s)}\right)^{k_s/2} w_2(n)^{(k_s-k_j)/2}, \tag{A.4}$$

and thus

$$\lim_{n\to\infty} C_{js} = \begin{cases} 0 & \text{if } k_j > k_s \\ 1 & \text{if } k_j = k_s \\ \infty & \text{if } k_j < k_s. \end{cases} \tag{A.5}$$

On the other hand, the limiting behaviour of the last factor in (2.16), which we denote by D_{js}, depends on whether M_s is nested within M_j. We therefore consider the following three situations:

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

Applying (A.3) we obtain

$$\operatorname*{plim}_{n\to\infty} D_{js} = \lim_{n\to\infty} \left(\frac{\sigma^2}{\sigma^2 + b_j}\right)^{(n-1)/2} = 0, \tag{A.6}$$

which, in combination with (A.5), leads directly 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.5) with (A.6) no longer leads directly to the limit of B_{js}. A natural **sufficient condition** leading to a zero limit for B_{js} is given in (3.4), 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.5) 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

$$D_{js} = \left(\frac{y' M_{X_s} y}{y' M_{X_j} y}\right)^{(n-1)/2} \left(1 + \frac{w_2(n)\{w_1(k_s)(A_s - 1) - w_1(k_j)(A_j - 1)\} + w_1(k_s)w_1(k_j)(A_s - A_j)}{\{w_2(n) + w_1(k_s)\}\{w_2(n) + w_1(k_j)A_j\}}\right)^{(n-1)/2}, \tag{A.7}$$

where

$$A_s = \frac{(y - \overline{y}\iota_n)'(y - \overline{y}\iota_n)}{y' M_{X_s} y} \text{ and } A_j = \frac{(y - \overline{y}\iota_n)'(y - \overline{y}\iota_n)}{y' M_{X_j} y},$$

it is immediate that the first factor in (A.7) 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 in (3.5) ensures a finite limit for the second factor. Alternatively, if (3.6) holds, the second factor is (A.7) is smaller than one. Thus, using the fact that C_{js} converges to zero, (3.5) and (3.6) each provide a sufficient condition for a zero limit of B_{js}.

A.2. Proof of results in Subsection 3.2 ($k_s \geq 1$)

A.2.1. plim(D_{js}) when M_s is nested within M_j.

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

$$\operatorname*{plim}_{n\to\infty} D_{js} = \lim_{n\to\infty} \left(\frac{\sigma^2 + \frac{g_{0s}}{g_{0s}+1} b}{\sigma^2 + \frac{g_{0j}}{g_{0j}+1} b} \right)^{(n-1)/2} = 0, \tag{A.8}$$

where b denotes the value b_j in (A.2) corresponding to $X_j = \iota_n$. This immediately leads to (3.7).

A.2.2. plim(D_{js}) when M_s is not nested within M_j.

Applying (A.1) $-$ (A.3) and assuming that

$$\frac{g_{0s}}{g_{0s}+1} b < \frac{g_{0j}}{g_{0j}+1} b + \frac{1}{g_{0j}+1} b_j, \tag{A.9}$$

where b corresponds to the model with just the intercept and b_j to X_j in (A.2), we obtain that

$$\operatorname*{plim}_{n\to\infty} D_{js} = \lim_{n\to\infty} \left(\frac{\sigma^2 + \frac{g_{0s}}{g_{0s}+1} b}{\sigma^2 + \frac{g_{0j}}{g_{0j}+1} b + \frac{1}{g_{0j}+1} b_j} \right)^{(n-1)/2} = 0. \tag{A.10}$$

Therefore, (A.9) ensures a zero limit for D_{js}, and we can deduce the results presented in Subsection 3.2 for this case.

REFERENCES

Akaike, H. (1981), "Likelihood of a Model and Information Criteria," *Journal of Econometrics*, 16, 3-14.

Amemiya, T. (1986), *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. and 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. (1980), "A Bayesian Analysis of Classical Hypothesis Testing," (with discussion) in *Bayesian Statistics*, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Valencia: University Press, 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*, Ser. A, 143, 383-430.

Chib, S. and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327-335.

Chow, G.C. (1981), "A Comparison of the Information and Posterior Probability Criteria for Model Selection," *Journal of Econometrics*, 16, 21-33.

Cornwell, C. and Trumbull, W.N. (1994), "Estimating the Economic Model of Crime With Panel Data," *Review of Economics and Statistics*, 76, 1994, 360-366.

Dawid, A.P. (1984), "Statistical Theory: The Prequential Approach," *Journal of the Royal Statistical Society*, Ser. A, 147, 278-292.

Dawid, A.P. (1986), "Probability Forecasting," in: *Encyclopedia of Statistical Sciences*, Vol. 7, eds. S. Kotz, N.L. Johnson, and C.B. Read, New York: Wiley, pp. 210-218.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty," (with discussion) *Journal of the Royal Statistical Society*, Ser. 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.

Freedman, D.A. (1983), "A Note on Screening Regressions," *The American Statistician*, 37, 152-155.

Geisser, S. and Eddy, W.F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153-160.

George, E.I. (1997), "Bayesian Model Selection," *Encyclopedia of Statistical Sciences*, New York: Wiley.

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

George, E.I. and McCulloch, R.E. (1997), "Approaches For Bayesian Variable Selection," *Statistica Sinica*, 7, 339-373.

Gelfand, A.E. and Dey, D.K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, Ser. B, 56, 501-514.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press, pp. 609-620.

Good, I.J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society*, Ser. 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. and Quinn, B.G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society*, Ser. B, 41, 190-195.

Kass, R.E. and Raftery, A.E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773-795.

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

Laud, P.W. and Ibrahim, J.G. (1995), "Predictive Model Selection", *Journal of the Royal Statistical Society*, Ser. B, 57, 247-262.

Laud, P.W. and 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*, New York: Wiley.

Lee, H. (1996), "Model Selection for Consumer Loan Application Data," mimeo, Carnegie Mellon University.

Madigan, D., Gavrin, J. and 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. and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215-232.

Min, C. and 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

Osiewalski, J. and Steel, M.F.J. (1993), "Regression Models Under Competing Covariance Structures: A Bayesian Perspective," *Annales d'Economie et de Statistique*, 32, 65-79.

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison," (with discussion) *Journal of the Royal Statistical Society*, Ser. B, 57, 99-138.

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 *Bayesian Statistics 2*, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, New York: Elsevier, 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 *Bayesian Statistics 5*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press, 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. and Hoeting, J.A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179-191.

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

Richard, J.F. and 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, A.F.M. and Spiegelhalter, D.J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 47, 213-220.

Vandaele, W. (1978), "Participation in Illegitimate Activities; Ehrlich Revisited," in *Deterrence and Incapacitation*, eds. A. Blumstein, J. Cohen and D. Nagin, Washington D.C.: National Academy of Sciences Press, pp. 270-335.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, eds. P.K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233-243.

Zellner, A. and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," (with discussion) in *Bayesian Statistics*, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Valencia: University Press, pp. 585-603.