Statistical Science
2015, Vol. 30, No. 2, 228–241
DOI: 10.1214/14-STS508

Posterior Model Consistency in Variable Selection as the Model Dimension Grows

Elías Moreno, Javier Girón and George Casella

Abstract. Most of the consistency analyses of Bayesian procedures for variable selection in regression refer to pairwise consistency, that is, consistency of Bayes factors. However, variable selection in regression is carried out in a given class of regression models where a natural variable selector is the posterior probability of the models.

In this paper we analyze the consistency of the posterior model probabilities when the number of potential regressors grows as the sample size grows. The novelty in the posterior model consistency is that it depends not only on the priors for the model parameters through the Bayes factor, but also on the model priors, so that it is a useful tool for choosing priors for both models and model parameters.

We have found that some classes of priors typically used in variable selection yield posterior model inconsistency, while mixtures of these priors improve this undesirable behavior.

For moderate sample sizes, we evaluate Bayesian pairwise variable selection procedures by comparing their frequentist Type I and II error probabilities. This provides valuable information to discriminate between the priors for the model parameters commonly used for variable selection.

Key words and phrases: Bayes factors, Bernoulli model priors, g-priors, hierarchical uniform model prior, intrinsic priors, posterior model consistency, rate of growth of the number of regressors, variable selection.

1. INTRODUCTION

In some applications of regression models to complex problems, for instance, in genomic, clustering, change points detection, etc., the dimension of the parameter space of the sampling models is either very large or grows with the sample size. The question we address here is whether consistency of the Bayesian variable selection approach still holds in this setting. A partial answer to this question was given in Moreno, Girón and Casella (2010), where consistency of the Bayes factors (pairwise consistency) when the number of regressors k increases with rate $k = O(n^b)$, $b \leq 1$, was considered. It was there proved that any pair of nested regression models for which the Bayes factor has an asymptotic approximation equivalent to the BIC (Schwarz, 1978) is consistent for $b < 1$ but it is not for $b = 1$. Note that the BIC is a valid approximation for a wide class of prior distributions on the model parameters. It was also seen that the Bayes factor for the intrinsic priors considerably improves the BIC behavior for small or moderate sample sizes (Casella et al., 2009).

Nevertheless, variable selection in regression is a model selection problem in a class $\mathfrak{M}$ of 2^k normal regression models, and we wonder if the Bayes factor consistency when $k = O(n^b)$, $b \leq 1$, can be extended to posterior model consistency in the class of models $\mathfrak{M}$. The use of the posterior model probabilities as a variable selector procedure implies that variable selection is understood as a decision problem where the decision space $\mathfrak{D}$ and the space of states of nature $\mathfrak{M}$ are

Elías Moreno is Professor of Statistics, Department of Statistics, University of Granada, 18071 Granada, Spain (e-mail: emoreno@ugr.es). Javier Girón is Professor of Statistics, Departament of Statistics, University of Málaga, 29016 Málaga, Spain (e-mail: fj_giron@uma.es). George Casella was Distinguished Professor, University of Florida, Florida, USA, deceased, June 17, 2012.

the same. Assuming a 0–1 loss function on the product space $\mathfrak{D} \times \mathfrak{M}$, the optimal decision is that of choosing the model with the highest posterior probability; other loss functions can indeed be used; see, for instance, the review paper by Clyde and George (2004).

Posterior model consistency in $\mathfrak{M}$ is understood as the convergence to one, in probability, of the sequence of the posterior probabilities of the true model. We are considering the true model to be the one from which the observations are drawn. We note that the frequentist and Bayesian consistency notions do not necessarily coincide. For instance, Shao (1997) defines a true model to be the submodel minimizing the average squared prediction error, and consistency of a model selection procedure means that the selected model converges in probability to this model.

From the necessary and sufficient conditions we give to achieve posterior model consistency it follows that Bayes factor consistency does not necessarily yield posterior model consistency. This was already pointed out by Johnson and Rossell (2012). Further, posterior model consistency of a Bayesian procedure in $\mathfrak{M}$ depends on the Bayes factor, the prior over the class of models $\mathfrak{M}$ and the rate of growth of k, and thus it has to be studied in a case-by-case basis.

The Bayes factors we review here are those obtained using the intrinsic priors on the model parameters (Berger and Pericchi, 1996; Moreno, 1997; Moreno, Bertolino and Racugno, 1998) and a couple of versions of the Zellner's g-priors (Zellner and Siow, 1980; Zellner, 1986). These versions include the g-priors with $g = n$ and the prior obtained as a mixture of g-priors with respect to the InverseGamma$(g|1/2, n/2)$. This latter prior was recommended by Zellner and Siow (1980) and considered in Liang et al. (2008) and Scott and Berger (2010), among others. As we will see, these Bayes factors exhibit different dimension corrections, that suggest a different behavior for moderate sample sizes, a point that we also explore here.

The priors over the set of models we review are the independent Bernoulli parametric class $\{\pi(M|\theta), 0 < \theta < 1\}$ introduced by George and McCulloch (1993) and a specific mixture of these priors which we refer to as the hierarchical uniform model prior $\pi^{\mathrm{HU}}(M)$. This latter prior is a particular case of a set of hierarchically uniform priors considered by George and McCulloch (1993), who argued that "*one may wish to weight more according the model size*."

Related posterior model consistency for variable selection for homoscedastic high-dimensional regression models was analyzed by Johnson and Rossell (2012). They considered Bayes factors for nonlocal priors on the regression parameters, an inverse gamma for the common variance errors, and models priors such that $\pi(M_t)/\pi(M) > \varepsilon > 0$ for any $M \in \mathfrak{M}$, where M_t, the true model, is a fixed model. We note that the Bernoulli class of model priors and the hierarchical uniform model prior $\pi^{\mathrm{HU}}(M)$ are excluded from their analysis. Further, the rate of growth of the number of regressors does not play a relevant role for the posterior model consistency of their Bayesian models, while for the Bayesian models considered here it does.

1.1 Notation

Let Y represent an observable random variable and $X_1, \ldots, X_k$ a potential set of explanatory regressors related through the normal linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon_k, \quad \varepsilon_k \sim N(0, \sigma_k^2),$$

where the vector of regression coefficients $\boldsymbol{\beta}_{k+1} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and the variance error σ_k^2 are unknown. Let $(\mathbf{y}, \mathbf{X})$ be the data set, where $\mathbf{y}$ is a vector of n independent observations of Y and $\mathbf{X}$ a $n \times (k+1)$ design matrix of full rank. This full sampling normal model $N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}_{k+1}, \sigma_k^2 \mathbf{I}_n)$ is denoted as M_k and the simplest intercept only normal model $N_n(\mathbf{y}|\beta_0 \mathbf{1}_n, \sigma_0^2 \mathbf{I}_n)$ as M_0. We remark that the regression coefficients change across models, although for simplicity we use the same alphabetical notation.

It is convenient to split the class $\mathfrak{M}$ of regression models involved in variable selection as follows. By $\mathfrak{M}_j$ we denote the class of models with j regressors, $0 \le j \le k$, the number of which is $\binom{k}{j}$, and by M_j we denote a generic model in $\mathfrak{M}_j$ with sampling density $N_n(\mathbf{y}|\mathbf{X}_{j+1}\boldsymbol{\beta}_{j+1}, \sigma_j^2 \mathbf{I}_n)$, where $\boldsymbol{\beta}_{j+1} = (\beta_0, \beta_1, \ldots, \beta_j)'$ is the unknown vector of regression coefficients, $\mathbf{X}_{j+1}$ a $n \times (j+1)$ submatrix of $\mathbf{X}$ and σ_j^2 the unknown variance error. Therefore, $\mathfrak{M} = \bigcup_{j=0}^{k} \mathfrak{M}_j$. The developments in the paper will be clear using this somewhat ambiguous, but simpler, notation.

1.2 Summary

We find that when k grows with n, the intrinsic priors for model parameters are preferred to either the g-prior for $g = n$ or the mixtures of g-priors, and the hierarchical uniform model prior is preferred to the Bernoulli model prior for any fixed value of the hyperparameter $\theta \in (0, 1)$.

The rest of the paper is organized as follows. In Section 2 we give necessary and sufficient conditions to achieve posterior model consistency. In Section 3 we give asymptotic approximations to the Bayes factors for the g-priors with $g = n$, for the mixture of g-priors and for the intrinsic priors over the model parameters, for $k = O(n^b)$, $0 \le b \le 1$. In Section 4 posterior model consistency for the Bayesian procedures is presented. Section 5 contains a sampling evaluation of the three Bayes factors for moderate sample sizes. A summary of the conclusions is given in Section 6, and the Appendix contains the proofs of most of the results.

2. POSTERIOR MODEL CONSISTENCY

Given a data set $(\mathbf{y}, \mathbf{X})$ coming from a linear model in $\mathfrak{M}$, and the priors for the models and model parameters $\{\pi(\boldsymbol{\beta}_{j+1}, \sigma_j | M_j), \pi(M_j),\ M_j \in \mathfrak{M}_j,\ j = 0, 1, \ldots, k\}$, the posterior probability of a generic model M_j can be written as

$$\text{(1)} \quad \Pr(M_j|\mathbf{y}, \mathbf{X}) = \frac{B_{j0}(\mathbf{y}, \mathbf{X})\pi(M_j)}{\sum_{i=0}^{k} \sum_{M_i \in \mathfrak{M}_i} B_{i0}(\mathbf{y}, \mathbf{X})\pi(M_i)},$$

where $B_{j0}(\mathbf{y}, \mathbf{X})$ denotes the Bayes factor for comparing models M_j and M_0, which is given by

$$\begin{aligned} &B_{j0}(\mathbf{y}, \mathbf{X}) \\ &\quad = \biggl(\int N_n(\mathbf{y}|\mathbf{X}_{j+1}\boldsymbol{\beta}_{j+1}, \sigma_j^2 \mathbf{I}_n) \\ &\qquad\qquad \cdot \pi(\boldsymbol{\beta}_{j+1}, \sigma_j | M_j)\, d\boldsymbol{\beta}_j\, d\sigma_j\biggr) \\ &\qquad \Big/ \biggl(\int N_n(\mathbf{y}|\beta_0, \sigma_0^2 \mathbf{I}_n)\pi(\beta_0, \sigma_0|M_0)\, d\alpha_0\, d\sigma_0\biggr). \end{aligned}$$

The advantage of the posterior model probability in expression (1) is that it only involves Bayes factors for nested models. The variable selection procedure that uses this posterior model probability as model selector is called encompassing from below variable selection (Girón et al., 2006). We may also use the encompassing from above approach in which all the Bayes factors considered are of the form $B_{jk}(\mathbf{y}, \mathbf{X})$ (Casella and Moreno, 2006). Both methods give similar results, and in this paper we will consider the encompassing from below approach.

DEFINITION. Posterior model consistency when sampling from model M_t holds if the limit in probability $[M_t]$ of the random variables $\{\Pr(M_j|\mathbf{y}, \mathbf{X}), M_j \in \mathfrak{M}\}$ is such that

$$\lim_{n\to\infty} \Pr(M_j|\mathbf{y}, \mathbf{X}) = \begin{cases} 1, & \text{if } j = t, \\ 0, & \text{if } j \neq t, \end{cases} \qquad [M_t].$$

A necessary and sufficient condition to achieve posterior model consistency when sampling from M_t is given in the next theorem.

THEOREM 1. *When sampling from $M_t \in \mathfrak{M}_t$, posterior model consistency holds if and only if the equality*

$$\text{(A)} \quad \lim_{n\to\infty} \sum_{j=0}^{k} \sum_{\substack{M_j \in \mathfrak{M}_j \\ M_j \neq M_t}} \frac{B_{j0}(\mathbf{y}, \mathbf{X})}{B_{t0}(\mathbf{y}, \mathbf{X})} \frac{\pi(M_j)}{\pi(M_t)} = 0, \quad [M_t],$$

holds.

PROOF. The assertion follows from expression (1). □

Theorem 1 implies that, if the Bayes factor $B_{t0}(\mathbf{y}, \mathbf{X})$ is inconsistent under M_t, then posterior model consistency under M_t does not hold. We note that when k is bounded, posterior model consistency holds for virtually any prior over the models (Casella et al., 2009). However, when $k = O(n^b)$, $0 < b \le 1$, it is apparent from (A) that posterior model consistency crucially depends on the rate of convergence under M_t to zero of the ratio $[B_{j0}(\mathbf{y}, \mathbf{X})\pi(M_j)]/[B_{t0}(\mathbf{y}, \mathbf{X})\pi(M_t)]$.

Under the null model M_0, the necessary and sufficient condition (A) reduces to

$$\text{(B)} \quad \lim_{n\to\infty} \sum_{j=1}^{k} \sum_{M_j \in \mathfrak{M}_j} B_{j0}(\mathbf{y}, \mathbf{X}) \frac{\pi(M_j)}{\pi(M_0)} = 0, \quad [M_0],$$

and it follows that if for some M_i the Bayes factor $B_{i0}(\mathbf{y}, \mathbf{X})$ is inconsistent under M_0, then posterior model consistency under M_0 does not hold. It is clear that, when $k = O(n^b)$, $0 < b \le 1$, the rate of convergence to zero of $B_{j0}(\mathbf{y}, \mathbf{X})\pi(M_j)$ determines the posterior model consistency.

Thus, from Theorem 1 it is clear that when k increases as the sample size n increases, posterior model consistency is a more stringent notion than that of the Bayes factor consistency. Furthermore, posterior model consistency depends on the specific Bayes factors B_{j0} and the prior on the class of models $\mathfrak{M}$, and, consequently, it has to be established in a case-by-case basis.

3. PRIORS AND BAYES FACTORS FOR VARIABLE SELECTION

In this section we present priors for the parameters of the models and priors over the class of models which are commonly used in variable selection. We give formulae for the Bayes factors and their asymptotic approximations when sampling from an arbitrary but fixed model M_t and rate of growth $k = O(n^b)$ for $0 \le b \le 1$.

3.1 Intrinsic Priors for Model Parameters

The intrinsic priors were introduced to justify the intrinsic Bayes factor (Berger and Pericchi, 1996). The original conditions defining the intrinsic priors given by Berger and Pericchi (1996) render a class of intrinsic priors (Moreno, 1997), and a limiting procedure for choosing a specific pair of intrinsic priors for model selection was proposed in Moreno, Bertolino and Racugno (1998). This procedure is based on the additional requirement that the intrinsic priors derived from improper priors, which are not necessarily proper, are a limit of proper intrinsic priors.

Bayes factors for intrinsic priors were used for variable selection in regression in Moreno and Girón (2005), Casella and Moreno (2006), Girón et al. (2006), Leon-Novelo, Moreno and Casella (2012), Consonni, Forster and La Rocca (2015), among others, and this variable selection procedure improves upon the Schwarz approximation for finite sample sizes (Casella et al., 2009) and asymptotically for high-dimensional regression models (Moreno, Girón and Casella, 2010).

The standard intrinsic method for comparing the null model M_0 versus the alternative M_j, starting from the improper reference prior for the parameters of the models M_0 and M_j, provides the proper intrinsic prior for the parameters $(\boldsymbol{\beta}_{j+1}, \sigma_j)$, conditional on a null point (α_0, σ_0), as

$$\begin{aligned}\pi^I(\boldsymbol{\beta}_{j+1}, \sigma_j|\alpha_0, \sigma_0) \\ = N_{j+1}\big(\boldsymbol{\beta}_{j+1}|\tilde{\boldsymbol{\alpha}}_0, (\sigma_j^2+\sigma_0^2)\mathbf{W}_{j+1}^{-1}\big)\,\mathrm{HC}^+(\sigma_j|\sigma_0),\end{aligned}$$

where $\tilde{\boldsymbol{\alpha}}_0 = (\alpha_0, \mathbf{0}_j')'$, $\mathbf{W}_{j+1}^{-1} = \frac{n}{j+2}(\mathbf{X}_{j+1}'\mathbf{X}_{j+1})^{-1}$, and

$$\mathrm{HC}^+(\sigma_j|\sigma_0) = \frac{2}{\pi}\frac{\sigma_0}{\sigma_j^2+\sigma_0^2}$$

is the half Cauchy distribution on R^+ with location parameter 0 and scale σ_0. The unconditional intrinsic prior with respect to the reference prior $\pi^N(\alpha_0, \sigma_0) = c_0/\sigma_0$ is then given by

$$\begin{aligned}\pi^I(\boldsymbol{\beta}_{j+1}, \sigma_j) \\ = \int \pi^I(\boldsymbol{\beta}_{j+1}, \sigma_j|\alpha_0, \sigma_0)\pi^N(\alpha_0, \sigma_0)\, d\alpha_0\, d\sigma_0.\end{aligned}$$

For comparing model M_0 versus M_j the intrinsic priors are the pair $(\pi^N(\alpha_0, \sigma_0), \pi^I(\boldsymbol{\beta}_{j+1}, \sigma_j))$. We note that $\pi^I(\boldsymbol{\beta}_{j+1}, \sigma_j)$ depends on the arbitrary constant c_0 that cancels out in the Bayes factor $B_{j0}(\mathbf{y}, \mathbf{X})$, and hence no tuning hyperparameters have to be adjusted. Thus, the Bayes factor for intrinsic priors are automatically constructed from the sampling models and the reference priors.

3.2 Zellner's g-Priors for Model Parameters

For variable selection with the g-priors we also use the encompassing from below approach (the encompassing from above version is given in Scott and Berger, 2010). A basic assumption on the regression models for constructing the g-priors is that the intercept and the variance error are common parameters to all models, which reduces the number of parameters involved when comparing M_j versus M_0 from $j+4$ to $j+2$. According to this restriction, the regression parameters of a generic model M_j will be denoted as $(\beta_0, \boldsymbol{\beta}_j)' = (\beta_0, \beta_1, \dots, \beta_j)'$ and the variance error as σ^2, where β_0 and σ are common to all models.

For a sample $(\mathbf{y}, \mathbf{X})$, most references to g-priors in the variable selection literature (Berger and Pericchi, 2001; Clyde and George, 2000; George and Foster, 2000; Fernández, Ley and Steel, 2001; Hansen and Yu, 2001; Liang et al., 2008, among others) refer to them as the pair $(\pi^N(\beta_0, \sigma), \pi^g(\boldsymbol{\beta}_j|\sigma))$, where

$$\pi^N(\beta_0, \sigma) = \frac{c_0}{\sigma}1_{R\times R^+}(\beta_0, \sigma)$$

is the reference prior, and

$$\pi(\boldsymbol{\beta}_j|\sigma, g) = N_j\big(\boldsymbol{\beta}_j|\mathbf{0}_j, g\sigma^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}\big),$$

g being an unknown positive hyperparameter, and $\mathbf{X}_j$ the matrix of dimensions $n \times j$ resulting from suppressing the first column in the design matrix $\mathbf{X}_{j+1}$ of the original formulation of the regression model M_j.

The conjugate property of these priors makes the expression of the Bayes factor quite simple, and it is well known that the hyperparameter g plays an important role in the behavior of the Bayes factor. Several values for g have been suggested, although none of them satisfies all the reasonable requirements (Berger and Pericchi, 2001; Clyde and George, 2004; Clyde, Parmigiani and Vidakovic, 1998; Fernández, Ley and Steel, 2001; George and Foster, 2000; Hansen and Yu, 2001; Liang et al., 2008). For instance, large g values induce the Lindley–Bartlett paradox (Bartlett, 1957), and a fixed value for g induces inconsistency, which can be corrected if g were dependent on n.

We consider two versions of the g-prior. The first is the one obtained for $g = n$, which is justified on the ground that it provides a consistent Bayes factor, and it is a "unit information prior" (Kass and Wasserman, 1995). The second g-prior version was derived for avoiding an incoherent property of the g-prior detected by Berger and Pericchi (2001): the Bayes factor for comparing M_j versus M_0 for the g-prior does not tend to infinity as the coefficient of determination of M_j tends to one. A way to avoid this behavior is to integrate the conditional g-priors $\{\pi(\boldsymbol{\beta}_j|\sigma, g), g > 0\}$ to obtain the mixture of g-priors

$$\pi^{\mathrm{Mix}}(\boldsymbol{\beta}_j|\sigma) = \int_0^\infty \pi(\boldsymbol{\beta}_j|\sigma, g)\pi(g)\, dg,$$

where

$$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} \exp\left(-\frac{n}{2g}\right).$$

This mixture has been considered by some other authors, including Clyde and George (2004), Liang et al. (2008) and Scott and Berger (2010).

3.3 Priors for Models

Since $\mathfrak{M}$ is a discrete space, a natural default prior over it is the uniform prior, but, as we will see, it is not a good prior when $k = O(n^b)$, $1/2 \le b \le 1$. A generalization of the uniform prior is the parametric independent Bernoulli prior class (George and McCulloch, 1993), for which the probability of a generic model M_j containing j out of k regressors, $j \le k$, is given by

$$\pi(M_j|\theta) = \theta^j (1-\theta)^{k-j}, \quad 0 \le \theta \le 1,$$

where θ is an unknown hyperparameter, the meaning of which is the probability of inclusion of a regressor in the model. The prior $\pi(M_j|\theta)$ assigns the same probability to models with the same dimension, and, in particular, for $\theta = 1/2$ the uniform prior is obtained.

If we assume a uniform distribution for θ, the unconditional probability of model M_j is given by

$$\pi^{\mathrm{HU}}(M_j) = \int_0^1 \theta^j (1-\theta)^{k-j}\, d\theta = \binom{k}{j}^{-1} \frac{1}{k+1}.$$

If we decompose this probability as

$$\pi^{\mathrm{HU}}(M_j) = \pi^{\mathrm{HU}}(M_j|\mathfrak{M}_j)\pi^{\mathrm{HU}}(\mathfrak{M}_j),$$

it follows that the model prior distribution, conditional on the class $\mathfrak{M}_j$, is uniform, and the marginal over the classes $\{\mathfrak{M}_j, j = 0, 1, \dots, k\}$ is also uniform. Then, it seems appropriate to call to this prior the hierarchical uniform prior.

We will see that the variable selection procedure that uses the hierarchical prior $\pi^{\mathrm{HU}}(M_j)$ outperforms the behavior of the one using the prior $\pi(M_j|\theta)$, for any value of θ.

3.4 Bayes Factors

For the data $(\mathbf{y}, \mathbf{X})$, it can easily be seen that the Bayes factor for comparing M_j versus M_0 for the g-prior with $g = n$ is given by

$$B_{j0}^{g=n}(\mathbf{y}, \mathbf{X}) = \frac{(1+n)^{(n-j-1)/2}}{(1+n\mathcal{B}_{j0})^{(n-1)/2}}, \tag{2}$$

for the mixture of g-priors by

$$B_{j0}^{\mathrm{Mix}}(\mathbf{y}, \mathbf{X}) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} \cdot \int_0^\infty \frac{(1+g)^{(n-j-1)/2}}{(1+g\mathcal{B}_{j0})^{(n-1)/2}} g^{-3/2} \exp\left(-\frac{n}{2g}\right) dg, \tag{3}$$

and for the intrinsic priors by

$$B_{j0}^{\mathrm{IP}}(\mathbf{y}, \mathbf{X}) = \frac{2}{\pi}(j+2)^{j/2} \cdot \int_0^{\pi/2} \frac{\sin^j \varphi (n + (j+2)\sin^2\varphi)^{(n-j-1)/2}}{(n\mathcal{B}_{j0} + (j+2)\sin^2\varphi)^{(n-1)/2}}\, d\varphi. \tag{4}$$

The integrals on $(0, \infty)$ in (3) and on $(0, \pi/2)$ in (4) do not have explicit expressions but need numerical integration.

We note that all these Bayes factors depend on the data through the statistic $\mathcal{B}_{j0}$, which is the ratio of the square sum of the residuals of models M_j and M_0, that is,

$$\mathcal{B}_{j0} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H}_j)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - (1/n)\mathbf{1}_n\mathbf{1}_n')\mathbf{y}}, \tag{5}$$

where $\mathbf{H}_j$ is the hat matrix associated to $\mathbf{X}_j$.

We observe that each Bayes factor exhibits a different dimension correction, and this suggests that for small or moderate samples sizes their behavior might be different, a point that we later explore in Section 5.

For large sample sizes n useful analytic approximations to the above Bayes factors are given in the next lemma.

LEMMA 1. *For large sample sizes* n, $k = O(n^b)$ *and* $0 \le b \le 1$, *the following approximations hold for any* $j \le k$:

(i)

$$B_{j0}^{g=n} \approx \begin{cases} n^{-j/2}\mathcal{B}_{j0}^{-n/2} \exp\left\{\frac{1}{2}\left(1 - \frac{1}{\mathcal{B}_{j0}}\right)\right\}, \\ \quad \textit{for } b < 1, \\ n^{-j/2}\mathcal{B}_{j0}^{-n/2} \exp\left\{\frac{1}{2}\left(1 - \frac{1}{\mathcal{B}_{j0}} - \frac{j}{n}\right)\right\}, \\ \quad \textit{for } b = 1, \end{cases} \tag{6}$$

(ii)

$$
(7)\quad B_{j0}^{\mathrm{Mix}} \approx \begin{cases} \left(\dfrac{n}{2}\right)^{-j/2} \mathcal{B}_{j0}^{-(n-j-2)/2} \dfrac{\Gamma((j+1)/2)}{\Gamma(1/2)}, \\ \quad \textit{for } b < 1, \\ \left(\dfrac{n}{2}\right)^{-j/2} \mathcal{B}_{j0}^{-(n-j-2)/2} \\ \quad \cdot \left(1 + \dfrac{j}{n}\mathcal{B}_{j0}\right)^{-(j+1)/2} \\ \quad \cdot \dfrac{\Gamma((j+1)/2)}{\Gamma(1/2)}, \\ \quad \textit{for } b = 1, \end{cases}
$$

(iii)

$$
(8)\quad B_{j0}^{\mathrm{IP}} \approx \begin{cases} \left(\dfrac{n}{j+2}\right)^{-j/2} \mathcal{B}_{j0}^{-(n-1)/2} \\ \quad \cdot \exp\left\{\dfrac{j+2}{2}\left(1 - \dfrac{1}{\mathcal{B}_{j0}}\right)\right\}, \\ \quad \textit{for } b < 1, \\ \left(1 + \dfrac{n}{j+2}\right)^{(n-j-1)/2} \\ \quad \cdot \left(1 + \dfrac{n\mathcal{B}_{j0}}{j+2}\right)^{-(n-1)/2}, \\ \quad \textit{for } b = 1. \end{cases}
$$

PROOF. See Appendix A. □

The next theorem summarizes the fact that the three Bayes factors have an equivalent expression for large samples sizes n and a bounded potential number of regressors k. Further, this expression is the one obtained by Schwarz (1978).

THEOREM 2. *When k is bounded, then, for large sample sizes n, the Bayes factors in (2), (3) and (4) are equivalent to the Schwarz approximation, that is,*

$$
B_{j0}^{g=n} \approx B_{j0}^{\mathrm{Mix}} \approx B_{j0}^{\mathrm{IP}} \approx \exp\left(-\frac{j}{2}\log n - \frac{n}{2}\log \mathcal{B}_{j0}\right).
$$

PROOF. The proof follows from Lemma 1 and some algebraic manipulations. □

Theorem 2 implies that for low-dimensional regular models, any Bayes factor is consistent, as the Schwarz approximation guarantees Bayes factor consistency. In this setting, for any positive model prior, posterior model consistency under an arbitrary model M_t also holds.

However, for high-dimensional models the Schwarz approximation does not necessarily guarantee either the Bayes factor consistency or the posterior model consistency. Other approximating forms than that of Schwarz appear in this latter setting.

3.5 Asymptotic Approximations to the Bayes Factors

The Bayes factor approximations in (6), (7) and (8) depend on the random sequence $\{\mathcal{B}_{j0}, n \geq 1\}$ given in (5). In this section we go a step forward and use the asymptotic distribution of the statistic $\mathcal{B}_{j0}$ under an arbitrary but fixed model M_t to approximate the Bayes factors $B_{j0}^{g=n}$, B_{j0}^{Mix} and B_{j0}^{IP}.

We first note that the asymptotic distribution of $\mathcal{B}_{j0}$ under M_t, a doubly noncentral beta distribution, depends on the limit of the pseudo-distance between models defined as

$$
\delta_n(M_t, M_j) = \frac{1}{2\sigma_t^2}\boldsymbol{\beta}_t' \frac{\mathbf{X}_t'(\mathbf{I}_n - \mathbf{H}_j)\mathbf{X}_t}{n}\boldsymbol{\beta}_t.
$$

General properties of this pseudo-distance have been studied in Girón et al. (2010). This pseudo-distance $\delta_n(M_t, M_j)$ can be simplified as follows. We first write the covariance matrix of the joint set of covariates of the model M_t and M_j, the dimensions of which are $(t+j)\times(t+j)$, as

$$
\Sigma_{t+j}^{(n)} = \begin{pmatrix} S_{tt}^{(n)} & S_{tj}^{(n)} \\ S_{jt}^{(n)} & S_{jj}^{(n)} \end{pmatrix},
$$

where the matrices $S_{tt}^{(n)}$, $S_{jj}^{(n)}$ are definite positive. Let us consider the matrices $S_{t\cdot j}^{(n)} = S_{tt}^{(n)} - S_{tj}^{(n)} S_{jj}^{(n)-1} S_{jt}^{(n)}$, and $S_{t\cdot j} = \lim_{n\to\infty} S_{t\cdot j}^{(n)}$. Then, it can now be seen that $\delta_n(M_t, M_j)$ can be expressed as

$$
\delta_n(M_t, M_j) = \frac{1}{2\sigma_t^2}\boldsymbol{\beta}_t' S_{t\cdot j}^{(n)} \boldsymbol{\beta}_t.
$$

In what follows we denote $\delta^*(M_t, M_j) = \lim_{n\to\infty}\delta_n(M_t, M_j)$, and if there is no confusion, $\delta^*(M_t, M_j)$ and $\delta_n(M_t, M_j)$ will be simply written as δ_{tj}^* and δ_{tj}.

For any M_j, using the asymptotic distribution of $\mathcal{B}_{j0}$ under M_t, we can now provide asymptotic approximations in probability $[M_t]$ to the Bayes factors $B_{j0}^{g=n}$, B_{j0}^{Mix} and B_{j0}^{IP} that we summarize in Lemma 2.

LEMMA 2. *When sampling from a model M_t, the Bayes factors in (2), (3) and (4) for $j \leq k = O(n^b)$,*

$b \leq 1$, *can be approximated for large n as*

$$(9)\quad B_{j0}^{g=n} \approx \begin{cases} n^{-j/2}\left(\dfrac{1+\delta_{tj}^*}{1+\delta_{t0}^*}\right)^{-n/2} \\ \quad \cdot \exp\left(\dfrac{\delta_{tj}^*-\delta_{t0}^*}{2(1+\delta_{tj}^*)}\right), \\ \quad \textit{for } b<1, \\ n^{-j/2}\left(\dfrac{1+\delta_{tj}^*-j/n}{1+\delta_{t0}^*}\right)^{-n/2} \\ \quad \cdot \exp\left(\dfrac{\delta_{tj}^*-\delta_{t0}^*-j/n}{2(1+\delta_{tj}^*-j/n)}\right), \\ \quad \textit{for } b=1, \end{cases}$$

$$(10)\quad B_{j0}^{\mathrm{Mix}} \approx \begin{cases} \left(\dfrac{ne}{j+1}\right)^{-j/2}\left(\dfrac{1+\delta_{tj}^*}{1+\delta_{t0}^*}\right)^{-(n-j-2)/2}, \\ \quad \textit{for } b<1, \\ \left(\dfrac{ne}{j+1}\right)^{-j/2} \\ \quad \cdot \left(\dfrac{1+\delta_{tj}^*-j/n}{1+\delta_{t0}^*}\right)^{-(n-j-2)/2}, \\ \quad \textit{for } b=1, \end{cases}$$

and

$$(11)\quad B_{j0}^{\mathrm{IP}} \approx \begin{cases} \left(\dfrac{n}{j+2}\right)^{-j/2}\left(\dfrac{1+\delta_{tj}^*}{1+\delta_{t0}^*}\right)^{-(n-j)/2}, \\ \quad \textit{for } b<1, \\ \left(1+\dfrac{n}{j}\right)^{(n-j-1)/2} \\ \quad \cdot \left(\dfrac{(n/j)(1+\delta_{tj}^*)+\delta_{t0}^*}{1+\delta_{t0}^*}\right)^{-(n-1)/2}, \\ \quad \textit{for } b=1. \end{cases}$$

PROOF. The proof follows from Lemma 1 and the asymptotic distribution of the statistic $\mathcal{B}_{j0}$ under model M_t (Casella et al., 2009), and it is omitted. □

From Lemma 2 it follows that when sampling from the null model M_0, that is, when $M_t = M_0$, the asymptotic approximations (9), (10) and (11) notably simplify, as they only depend on n and the dimension j of the model, irrespective of the particular set of covariates. This means that, under the null model M_0, the above Bayes factors are asymptotically constant across models in the class $\mathfrak{M}_j$.

To prove some results on posterior model consistency when sampling from an alternative model M_t, we need to know for which models M_j in $\mathfrak{M}$ the pseudodistance $\delta^*(M_t, M_j)$ is zero. This result follows from Lemma 3.

LEMMA 3. (i) *For any model* M_j *such that* $\dim(M_j) < \dim(M_t)$, *we have that*

$$\delta^*(M_t, M_j) > 0.$$

(ii) *For any model* M_j *such that* $\dim(M_j) = \dim(M_t)$,

$$\delta^*(M_t, M_j) = \begin{cases} 0, & \textit{if } M_j = M_t, \\ >0, & \textit{if } M_j \neq M_t. \end{cases}$$

(iii) *For any model* M_j *such that* $\dim(M_j) > \dim(M_t)$,

$$\delta^*(M_t, M_j) = \begin{cases} 0, & \textit{if } M_t \textit{ is nested in } M_j, \\ >0, & \textit{otherwise}. \end{cases}$$

PROOF. We note that (a) if M_t and M_j do not have common covariates, then the matrix $\Sigma_{t+j} = \lim_{n\to\infty} \Sigma_{t+j}^{(n)}$ is positive definite, and hence $S_{t\cdot j}$ is positive definite, and (b) if M_t and M_j do have common covariates, then it can be seen that

$$S_{t\cdot j} = \begin{pmatrix} \mathbf{P} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix},$$

where $\mathbf{P}$ is a positive definite matrix of dimensions $\max\{0, \dim M_t - \dim M_j\}$. We observe that if either $\dim M_t = \dim M_j$ or M_t is nested in M_j, we have that $S_{t\cdot j} = \mathbf{O}$. The proof of Lemma 3 follows from (a) and (b) and the fact that all regression coefficients $\boldsymbol{\beta}_t$ in model M_t are different from zero. □

It is interesting to remark that for $b < 1$ and any M_j such that $\delta^*(M_t, M_j) > 0$, the rate of convergence in probability $[M_t]$ to zero of $B_{j0}^{g=n}$, B_{j0}^{Mix} and B_{j0}^{IP} for $M_t \neq M_0$ is exponentially fast, but the rate of convergence in probability $[M_0]$ to zero for $j \neq 0$ is only potentially fast. This is in line with the result for $b = 0$ obtained by Dawid (2011) (for discrete data see Consonni, Forster and La Rocca, 2015).

4. POSTERIOR MODEL CONSISTENCY FOR $k = O(n^b)$ AND $0 \leq b \leq 1$

Posterior model consistency results for the six Bayesian variable selection procedures defined by the Bayes factors $B_{j0}^{g=n}$, B_{j0}^{Mix}, B_{j0}^{IP}, the Bernoulli model prior $\pi(M_j|\theta)$ and the hierarchical uniform prior $\pi^{\mathrm{HU}}(M_j)$, when sampling from an arbitrary but fixed model M_t are summarized in Theorem 3. For simplicity, the posterior model consistency results for the case when sampling from model M_0 and from an alternative model M_t are not separated. However, we keep in mind that the rate of convergence of the posterior model probabilities when sampling from M_0 is different from the rate of convergence when sampling from $M_t \neq M_0$.

THEOREM 3. (i) *When sampling from M_t and $k = O(n^b)$, the Bayesian procedures for the Bayes factors $B_{j0}^{g=n}$, B_{j0}^{Mix}, B_{j0}^{IP}, and the Bernoulli model prior are posterior model consistent for $0 \le b < 1/2$ and posterior model inconsistent for $1/2 \le b \le 1$.*

(ii) *When sampling from M_t and $k = O(n^b)$, the Bayesian procedures for the Bayes factors $B_{j0}^{g=n}$, B_{j0}^{Mix} and B_{j0}^{IP} and the hierarchical uniform prior are posterior model consistent for $0 \le b \le 1$.*

PROOF. See Appendix B. □

It is interesting to observe that the Bernoulli prior $\pi(M_j|\theta)$, conditional on θ, induces a Binomial distribution on the classes $\mathfrak{M}_j$, which, in turn, by the change of variables $x = j/k$, induces a distribution on $x \in [0, 1]$ which converges in probability to a Dirac's delta on θ, as k tends to infinity. In other words, for large values of k the Bernoulli prior concentrates around models which have a proportion of covariates close to θ. Therefore, this apparently innocuous prior conveys too much prior information about the proportion of covariates of the models, and thus it makes the posterior model probabilities for $1/2 \le b \le 1$ inconsistent. This wrong asymptotic behavior is corrected by the hierarchical uniform prior.

5. SMALL SAMPLE COMPARISONS

Given a Bayes factor B_{j0} for the models $\{M_0, M_j\}$, the decision of choosing model M_j when $\Pr(M_j|B_{j0}) \ge 1/2$ is an optimal decision under $\Pr(M_0) = \Pr(M_j) = 1/2$ and a 0–1 loss function. We recall that for a uniform prior on the class of models $\mathfrak{M}$, to rank the models in the class according to their posterior model probabilities is equivalent to the ranking produced by the Bayes factor. In spite of this, a sampling analysis of the optimal statistical decision function has been long claimed [see, e.g., Fraser (2011) and discussions therein]. From expression (2), (3) and (4) it is apparent that the dimension correction of the Bayes factors for the g-prior with $g = n$, for the mixture of g-priors and for the intrinsic priors are different from each other. This suggests that their sampling behavior for small and moderate sample sizes might be different.

In this section we study the sampling properties of the posterior model probabilities for $\Pr(M_0) = \Pr(M_j) = 1/2$ and the Bayes factors $B_{j0}^{g=n}$, B_{j0}^{Mix} and B_{j0}^{IP}. We recall that the posterior probability $\Pr(M_j|\mathbf{y}, \mathbf{X})$ for any of these Bayes factors depends on the data $(\mathbf{y}, \mathbf{X})$ through the same statistic $\mathcal{B}_{j0}$, which

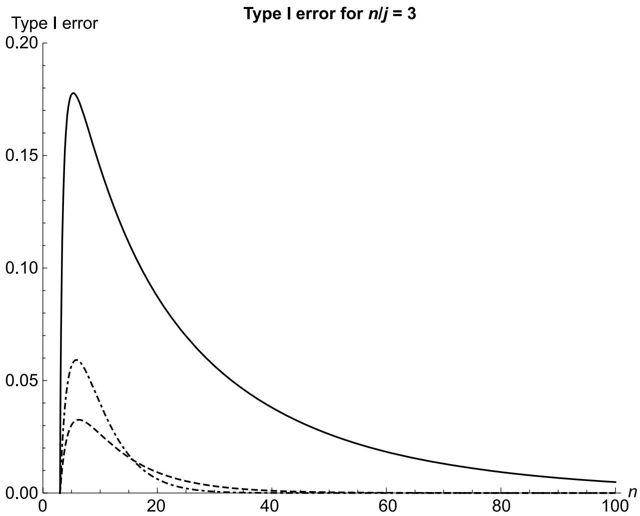

FIG. 1. *Type I error probabilities for the intrinsic procedure (continuous line), the g-prior with $g = n$ (dot-dashed) and the mixture of g-priors (dashed).*

takes values in the interval $(0, 1)$. Therefore, the critical regions for rejecting the null model M_0 for these Bayesian procedures are

$$R_{j0}^{(g=n)} = \{\mathcal{B}_{j0} : \Pr(M_j|B_{j0}^{g=n}) \ge 1/2\},$$
$$R_{j0}^{\mathrm{Mix}} = \{\mathcal{B}_{j0} : \Pr(M_j|B_{j0}^{\mathrm{Mix}}) \ge 1/2\}$$

and

$$R_{j0}^{\mathrm{IP}} = \{\mathcal{B}_{j0} : \Pr(M_j|B_{j0}^{\mathrm{IP}}) \ge 1/2\}.$$

These critical regions are in $(0, 1)$ and, since the posterior probabilities are monotone increasing functions of $\mathcal{B}_{j0}$, they are intervals. Using the distribution of the statistic $\mathcal{B}_{j0}$ under M_0 and M_j, we can compute the exact value of the Type I and II errors probabilities as a function of the model dimension j and the sample size n. Figure 1 shows the Type I error probabilities of the optimal decision rule associated to the regions $R_{j0}^{(g=n)}$, R_{j0}^{Mix} and R_{j0}^{IP} for $j = n/3$ and the sample size $n = 1, \dots, 100$.

From Figure 1 it follows that the Type I error probabilities of the procedures based on g-priors are very close to each other and smaller than that based on the intrinsic priors. We note that as n and j increase at the same pace, $n/j = 3$, the Type I error probabilities for the procedure based on g-priors go faster to zero than the procedure based on the intrinsic priors does.

In Figure 2 we display for $\delta_{j0} = 1$ and $j = n/3$ the power function of the above procedures as a function of the sample size $n = 1, \dots, 100$.

From Figure 2 we observe that the power of the procedure based on intrinsic priors is much larger than

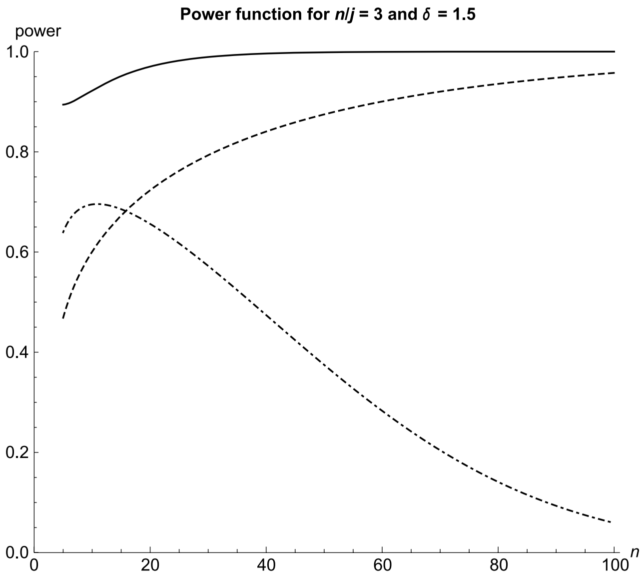

FIG. 2. *Power for the Bayes factor for intrinsic priors (continuous line), for g-priors for $g = n$ (dot-dashed) and for the mixture of g-priors (dashed).*

those based on the g-priors. This is the price the procedures based on g-priors pay for their very small Type I error probabilities. Further, the power for the intrinsic priors and the mixture of g-priors increases to one as the sample size n and the model dimension j increase at the same pace, that is, $n/j = r \geq 1$, but the power for the g-prior with $g = n$ increases as n increases up to a certain n and then decreases, which is a surprisingly unreasonable behavior. The explanation to the anomalous behavior of the Bayes factor for the $g = n$ is due to the inconsistency of this Bayes factor for any model M_j such that $j = O(n)$, a point that we discuss in Section 6 and summarize in Table 2.

On the other hand, we remark that as δ_{j0} increases, the power of the three procedures increases for any sample size n.

Figures 1 and 2 indicate how unbalanced are the Type I and II error probabilities of the Bayesian procedures based on g-priors compared with that based on the intrinsic priors. The practical implications of this analysis are that for moderate sample sizes the Bayesian procedures based on g-priors are strongly biased toward the null model.

6. CONCLUDING REMARKS

Variable selection in regression is a central problem in statistical inference and the aim of this paper has been to evaluate the sampling properties of Bayesian model selection procedures, a requirement long advocated by many statisticians. For some interesting applications the number of regressors is very large, and hence we assumed that the potential number of regressors k grows with n. We very soon realized that the variable selection takes place in a large class of models, and hence posterior model consistency seems to be the appropriate asymptotic property to be explored, a concept that depends on the priors over the models and the model parameters. Posterior model consistency for variable selection for three popular Bayes factors and two types of model priors has been explored, although the methodology we used can be applied to any other specific Bayes factor and model prior.

For low-dimensional normal regression models it is well known that virtually any Bayes factor has an asymptotic approximation which is equivalent to the Schwarz approximation, which assures consistency. However, for large-dimensional models more appropriate asymptotic approximations for the Bayes factors, such as those given in Lemma 2, are necessary for analyzing consistency.

Although we considered the independent Bernoulli class of model priors $\{\pi(M|\theta),\ \theta \in (0,1), M \in \mathfrak{M}\}$ and the hierarchical uniform prior $\pi^{\mathrm{HU}}(M)$, a mixture of $\pi(M|\theta)$ with respect to the uniform distribution on θ, the asymptotic results for the hierarchical uniform prior can be formally extended to other regular mixtures of Bernoulli model priors.

The conclusions on posterior model consistency we draw for the above Bayesian procedures when sampling from an arbitrary but fixed model M_t and for different rates of growth of k are summarized in Table 1.

Table 1 implies that when sampling from M_t, the Bayesian procedures for the Bayes factors $B_{j0}^{g=n}$, B_{t0}^{Mix} and B_{t0}^{IP} and the Bernoulli model prior are inconsistent for $1/2 \leq b \leq 1$, but for the hierarchical uniform model prior they are consistent for any $0 \leq b \leq 1$. Thus, a first conclusion is that the hierarchical uniform model prior

TABLE 1
Posterior model consistency when sampling from M_t, as a function of the Bayes factor, model prior and the rate of growth of $k = O(n^b)$

Model prior	$\boldsymbol{\pi(M\vert\theta)}$	$\boldsymbol{\pi^{\mathrm{HU}}(M)}$
Bayes factor	$B_{j0}^{g=n}$, B_{j0}^{Mix}, B_{j0}^{IP}	$B_{j0}^{g=n}$, B_{j0}^{Mix}, B_{j0}^{IP}
$0 \leq b < 1/2$	Consistent	
$1/2 \leq b \leq 1$	Inconsistent	Consistent

$\pi^{\mathrm{HU}}(M)$ outperforms the independent Bernoulli model prior $\pi(M|\theta)$.

We remark that the above results are valid when sampling from a fixed model M_t with finite dimension. The analysis of the infinite dimensional case is an open problem that deserves more efforts, as the Bayes factors are now not necessarily consistent (Moreno, Girón and Casella, 2010) and, consequently, the posterior model consistency results differ from those presented above. For instance, for $t = O(n)$, the Bayes factor $B_{t0}^{g=n}$ is such that

$$\lim_{n\to\infty} B_{t0}^{g=n} = 0, \quad [M_t],$$

so that it is inconsistent under any model $M_t \neq M_0$, and this implies that it is also posterior model inconsistent under $M_t \neq M_0$ for any model prior.

For the Bayes factors B_{t0}^{Mix} and B_{t0}^{IP} the situation is not so dramatic. The set of alternative models M_t for which inconsistency of B_{t0}^{Mix} holds is a small set of models around M_0 that satisfy the condition

$$\delta_{t0}^* < \delta_{\mathrm{mix}}(r) = \left(1 - \frac{1}{r}\right)(er)^{1/(r-1)} - 1,$$

where $r = n/t > 1$. Likewise, the set of alternative models M_t for which B_{t0}^{IP} is inconsistent is that given by the condition

$$\delta_{t0}^*(r) < \delta_{\mathrm{IP}}(r) = \frac{r-1}{(r+1)^{(r-1)/r}} - 1.$$

A summary of these results is given in Table 2.

From Table 2 we can draw the conclusion that the intrinsic priors and the mixture of g-priors are preferred to the g-prior for $g = n$.

We also note that $\delta_{\mathrm{IP}}(r) < \delta_{\mathrm{mix}}(r)$, so that the inconsistency region of the Bayes factor for the intrinsic priors is smaller than that for the mixture of g-priors. Further, for the case where $r = 1$ it can be shown that the Bayes factor B_{t0}^{Mix} is inconsistent for any alternative model M_t, while the Bayes factor B_{t0}^{IP} is inconsistent only for those M_t such that $\delta_{t0}^* < 1/\log 2 - 1$.

TABLE 2
Posterior model consistency when sampling from M_t, for $t = n/r$, $r > 1$, and $\pi(M_0) > 0$

Bayes factor	**Model prior**	**Posterior model consistency**
$B_{t0}^{g=n}$	$\pi(M_0) > 0$	Inconsistent under any M_t
B_{t0}^{Mix}	$\pi(M_0) > 0$	Inconsistent under M_t such that $\delta_{t0}^* < \delta_{\mathrm{mix}}(r)$
B_{t0}^{IP}	$\pi(M_0) > 0$	Inconsistent under M_t such that $\delta_{t0}^* < \delta_{\mathrm{IP}}(r)$

On the other hand, for small and moderate sample sizes, Figures 1 and 2 that we presented indicate that the behavior of the Bayes factors $B_{t0}^{g=n}$ and B_{j0}^{Mix} are strongly biased toward the null model, while the Bayes factor for the intrinsic priors B_{j0}^{IP} has more balanced Type I and II error probabilities.

Therefore, the overall conclusion from our analysis is that the intrinsic priors over the model parameters and the hierarchical uniform prior over the models are nowadays the priors to be recommended for variable selection in normal regression.

APPENDIX A: PROOF OF LEMMA 1

Part (i) is immediate and hence it is omitted. Part (ii) follows by first making the change of variables $y = \exp[-n/(2g)]$ in the integral in (3). The Jacobian of the inverse transformation is

$$J = \frac{dg}{dy} = \frac{n}{2y \log y^2}$$

and, thus, the integral in (3) now becomes

$$\int_0^1 \left(1 - \frac{n}{2\log y}\right)^{(n-j-1)/2} \left(1 - \frac{n\mathcal{B}_{j0}}{2\log y}\right)^{(-n+1)/2} \cdot \left(-\frac{n}{2\log y}\right)^{-3/2} y J \, dy.$$

The first factor in this integral can be approximated by

$$\left(1 - \frac{n}{2\log y}\right)^{(n-j-1)/2} \approx y^{-1+(j/n)} \left(-\frac{n}{2\log y}\right)^{(n-j-1)/2}$$

and the second by

$$\left(1 - \frac{n\mathcal{B}_{j0}}{2\log y}\right)^{(-n+1)/2} \approx \mathcal{B}_{j0}^{(1-n)/2} y^{1/\mathcal{B}_{j0}} \left(-\frac{n}{2\log y}\right)^{(1-n)/2}.$$

Plugging these approximations in the integral, and after some simplifications, we obtain that the original Bayes factor can be approximated as

$$B_{j0}^{\mathrm{Mix}}(\mathbf{y}, \mathbf{X}) \approx \frac{(n/2)^{-j/2}}{\Gamma(1/2)} \mathcal{B}_{j0}^{(-n+1)/2} \cdot \int_0^1 y^{(1/\mathcal{B}_{j0})-1+(j/n)} \left(-\frac{1}{\log y}\right)^{(1-j)/2} dy.$$

For any j and n, the integral in this expression has value

$$\int_0^1 y^{(1/\mathcal{B}_{j0})-1+(j/n)}\left(-\frac{1}{\log y}\right)^{(1-j)/2} dy = \mathcal{B}_{j0}^{(j+1)/2}\left(1+\frac{j}{n}\mathcal{B}_{j0}\right)^{-(j+1)/2}\Gamma\left(\frac{j+1}{2}\right),$$

and thus the approximation of the Bayes factor is

$$B_{j0}^{\mathrm{Mix}}(\mathbf{y},\mathbf{X}) \approx \left(\frac{n}{2}\right)^{-j/2}\mathcal{B}_{j0}^{-(n-j-2)/2}\left(1+\frac{j}{n}\mathcal{B}_{j0}\right)^{-(j+1)/2} \cdot \frac{\Gamma((j+1)/2)}{\Gamma(1/2)}.$$

If $b<1$, we have that

$$\lim_{n\to\infty}\left(1+\frac{j}{n}\mathcal{B}_{j0}\right)^{-(j+1)/2}=1,$$

and this proves the first part of (ii). If $b=1$, the proof follows suit directly from the expression of the approximation. This completes the proof of part (ii).

Part (iii) was proved in Girón et al. (2010) and hence it is omitted.

APPENDIX B: PROOF OF THEOREM 3

1. We first prove that condition (A) holds for the Bayes factor $B_{j0}^{g=n}$, the Bernoulli model prior $\pi(M_j|\theta)$ and $0\le b<12$, and that it does not hold for $1/2\le b\le 1$. For, under the Bernoulli prior we have that

$$(\mathrm{A}) = \sum_{j=0}^{k}\sum_{\substack{M_j\in\mathfrak{M}_j\\ M_j\neq M_t}} n^{(t-j)/2}(1+\delta_{tj}^*)^{-n/2} \cdot \exp\left\{\frac{1}{2}\left(\frac{\delta_{tj}^*-\delta_{t0}^*}{1+\delta_{tj}^*}\right)\right\}\left(\frac{\theta}{1-\theta}\right)^{j-t}.$$

From Lemma 3, the terms for $j\le t$ go to zero as n tends to infinity. For $j>t$ let us split the class $\mathfrak{M}_j$ as

$$\mathfrak{M}_j=\mathfrak{N}_j\cup(\mathfrak{M}_j-\mathfrak{N}_j),$$

where $\mathfrak{N}_j$ is the class of models M_j such that M_t is nested in M_j. From Lemma 3, it follows that $\delta_{tj}^*=0$ for any $M_j\in\mathfrak{N}_j$, and $\delta_{tj}^*>0$ for $M_j\in\mathfrak{M}_j-\mathfrak{N}_j$. Therefore, for large n the contribution of the models in $\mathfrak{M}_j-\mathfrak{N}_j$ to the sum in (A) tends to zero, and we then have for large n that

$$\begin{aligned}(\mathrm{A}) &\approx \sum_{j=t+1}^{k}\sum_{M_j\in\mathfrak{N}_j} n^{(t-j)/2} \cdot \exp\left\{\frac{1}{2}\left(\frac{\delta_{tj}^*-\delta_{t0}^*}{1+\delta_{tj}^*}\right)\right\}\left(\frac{\theta}{1-\theta}\right)^{j-t}\\
&\approx \sum_{i=1}^{k-t}\binom{k-t}{i}n^{-i/2}\left(\frac{\theta}{1-\theta}\right)^{i}\\
&=\left(-1+\left(1+\frac{\theta}{(1-\theta)n^{1/2}}\right)^{k-t}\right)\\
&\approx \exp\{n^{b-1/2}\}.\end{aligned}$$

Then, for large n, (A) is equivalent to $\exp\{n^{b-1/2}\}$ and this proves the assertions.

2. We now prove that for the Bayes factor $B_{j0}^{g=n}$ and the hierarchical uniform prior $\pi^{\mathrm{HU}}(M)$, condition (A) does hold for any $b\le 1$. Indeed, for large n, using again the decomposition $\mathfrak{M}_j=\mathfrak{N}_j\cup(\mathfrak{M}_j-\mathfrak{N}_j)$, the sum (A) can be approximated for large n as

$$\begin{aligned}(\mathrm{A}) &\approx \sum_{j=t+1}^{k}\sum_{M_j\in\mathfrak{N}_j} n^{(t-j)/2}\exp\left\{\frac{1}{2}\left(\frac{\delta_{tj}^*-\delta_{t0}^*}{1+\delta_{tj}^*}\right)\right\}\frac{\binom{k}{t}}{\binom{k}{j}}\\
&=\sum_{i=1}^{k-t}n^{-i/2}\frac{(t+i)!}{i!}\\
&<\sum_{i=1}^{\infty}n^{-i/2}\frac{(t+i)!}{i!}\\
&=t!\left(-1+(1-n^{-1/2})^{-t}\frac{n^{1/2}}{n^{1/2}-1}\right).\end{aligned}$$

The last expression tends to zero as n tends to infinity, and this proves the assertion.

3. Let us now consider the Bayes factor B_{j0}^{IP} and the Bernoulli prior. For simplicity we prove the assertion for $\theta=1/2$, as the proof for any θ follows the same line of reasoning. We first note that the contribution of the models in $\mathfrak{M}_j-\mathfrak{N}_j$ to the sum in (A) tends to zero as n tends to infinity. Thus, we have for large n that

$$(\mathrm{A})\approx(t+2)^{-t/2}\sum_{j=t+1}^{k}\left(\frac{n}{j+2}\right)^{-j/2}n^{t/2}\binom{k-t}{j-t},$$

which, after the change of variables $i=j-t$, adopts the form

$$(\mathrm{A})\approx(t+2)^{-t/2} \cdot \sum_{i=1}^{k-t}a_i\frac{(k-t)(k-t-1)\cdots(k-t-i+1)}{n^{i/2}},$$

where

$$a_i = \frac{(t+i+2)^{(t+i)/2}}{i!}.$$

It can be seen that the sequence $\{a_i\}$ increases as i increases for $i < i_0(t)$, where $i_0(t) \approx [1 + 1.65\sqrt{t}]$, and decreases for $i > i_0(t)$, and thus it is bounded by some function of t, say, $a(t)$. Thus, the sum in (A) is upper bounded as

$$\begin{aligned}(\mathrm{A}) \le {} & (t+2)^{-t/2} a(t) \\ & \cdot \sum_{i=1}^{k-t} \frac{(k-t)(k-t-1)\cdots(k-t-i+1)}{n^{i/2}},\end{aligned}$$

which, for $b < 1/2$, converges to 0 as n tends to infinity. A similar lower bound for (A) shows that for $b \ge 1/2$ the sum cannot converge to zero.

For the Bayes factor B_{j0}^{Mix} the proof of the posterior model consistency is similar and hence omitted.

4. We now prove that for B_{j0}^{IP} and $\pi^{\mathrm{HU}}(M_j)$ posterior consistency holds for $b < 1$. For large n we have that

$$\begin{aligned}(\mathrm{A}) \approx {} & (t+2)^{-t/2} \\ & \cdot \sum_{j=t+1}^{k} \left(\frac{n}{j+2}\right)^{-j/2} n^{t/2} \binom{k-t}{j-t} \frac{j!(k-j)!}{t!(k-t)!},\end{aligned}$$

which simplifies to

$$(\mathrm{A}) \approx (t+2)^{-t/2} \sum_{j=t+1}^{k} \left(\frac{n}{j+2}\right)^{-j/2} n^{t/2} \frac{j!}{(j-t)!}.$$

Making the change of variable $i = j - t$, the expression adopts the form

$$(\mathrm{A}) \approx (t+2)^{-t/2} \sum_{i=1}^{k-t} \frac{b_i}{n^{i/2}},$$

where

$$b_i = (t+i+2)^{(t+i)/2} \frac{(t+i)!}{i!}.$$

Every individual term $b_i/n^{i/2}$ in the sum converges to 0 as n tends to infinity, and for large values of i, the summands $b_i/n^{i/2}$ can be approximated by

$$\frac{e^{t/2+1}\, i^{(i+3t)/2}}{n^{i/2}}.$$

For every t, this function of i is decreasing for all $i < i_0$ and increasing for $i > i_0$, where i_0 is given by

$$i_0 = -\frac{3t}{W(-3et/n)} \approx n/e - 3t.$$

But as $k = O(n^b)$ with $b < 1$, this implies that the sequence $b_i/n^{i/2}$ is decreasing in i for all $i \le k$. Then, it follows that the sum is upper bounded as

$$\sum_{i=i_0+1}^{k-t} b_i/n^{i/2} \le \frac{b_1}{n^{1/2}} + (k-t)\frac{b_2}{n}.$$

For $k = O(n^b)$ with $b < 1$, the limit of the right-hand side of this equation is 0 when n tends to infinity, and hence posterior model consistency holds.

The proof of the consistency for the Bayes factor for the mixture of g-priors follows exactly the same pattern and it is therefore omitted.

5. For $b = 1$, the proof of the posterior model consistency for B_{j0}^{Mix} and $\pi^{\mathrm{HU}}(M_j)$ runs as follows. For large n, it follows that, under the alternative model M_t,

$$\begin{aligned}\frac{B_{j0}^{\mathrm{Mix}}}{B_{t0}^{\mathrm{Mix}}} \approx {} & \frac{((ne)/(j+1))^{-j/2}}{((ne)/(t+1))^{-t/2}} \\ & \cdot \frac{(1+\delta_{tj}^{*} - j/n)^{-(n-j-2)/2}}{(1-t/n)^{-(n-t-2)/2}}.\end{aligned}$$

The ratio $\pi^{\mathrm{HU}}(M_j)/\pi^{\mathrm{HU}}(M_t)$ of the model probabilities for the hierarchical uniform prior is

$$\frac{\pi^{\mathrm{HU}}(M_j)}{\pi^{\mathrm{HU}}(M_t)} = \frac{j!(k-j)!}{t!(k-t)!}.$$

Then, reasoning as before, for large n, the double sum (A) of Theorem 1, after some simplifications, can be approximated as

$$\begin{aligned}(\mathrm{A}) \approx {} & (t+1)^{-t/2} \\ & \cdot \sum_{j=t+1}^{k} \left(\frac{ne}{j+1}\right)^{-j/2} n^{t/2} \\ & \cdot \frac{j!}{(j-t)!}\left(1 - \frac{j}{n}\right)^{-(n-j-2)/2}.\end{aligned}$$

Making the change of variable $i = j - t$, some further simplifications on the factorials yield the approximating expression

$$\begin{aligned}(\mathrm{A}) \approx {} & \frac{e^{-3t/2}}{(t+1)^{t/2}} \sum_{i=1}^{k-t} \frac{e^{-i/2}\, i^{-(i+1/2)} (i+t)^{(3i+3t)/2}}{n^{i/2}} \\ & \cdot \left(1 - \frac{i+t}{n}\right)^{-(n-i-t-2)/2}.\end{aligned}$$

Letting $x = i/k$ and $s = n/k$, the sum in the preceding expression can be approximated, up to a constant, by the integral $\int_0^1 f_k(x|s,t)\,dx$, where

$$f_k(x|s,t) = k(kx)^{-kx-(1/2)}(ks)^{-(kx)/2} \cdot e^{-(1/2)kx}(kx+t)^{(3kx)/2+(3t)/2+1/2} \cdot \left(1-\frac{kx+t}{ks}\right)^{(1/2)(k(x-s)+t+2)}.$$

We now prove that $\lim_{k\to\infty}\int_0^1 f_k(x|s,t)\,dx = 0$ for any $t=0,1,2,\dots$ and $s\geq 1$.

For any k, t and $s\geq 1$, $f_k(x|s,t)>0$. For $t=0$, we have that $f_k(x|s,0)$ is a decreasing function of x for all k and $s\geq 1$, and such that $f_k(0|s,0)=k$. Further, $\lim_{k\to\infty} f_k(x|s,0)=0$ for all $x\in(0,1]$. For $t=1,2,\dots,$ even though $f_k(x|s,t)$ is not a decreasing function of x, except for large values of x, we have that $\lim_{x\to 0} f_k(x|s,t)=0$, and $\lim_{k\to\infty} f_k(x|s,0)=0$ for all $x\in(0,1]$.

Thus, for any t, the limit of $f_k(x|s,t)$ when k goes to infinity is given by

$$\lim_{k\to\infty} f_k(x|s,t) = \begin{cases}\infty, & \text{if } x=0,\\ 0, & \text{if } x\in(0,1],\end{cases}$$

and thus

$$\int_0^1 \lim_{k\to\infty} f_k(x|s,t)\,dx = 0.$$

On the other hand, $f_k(x|s,t)$ is a decreasing function of s and, therefore, $f_k(x|s,t)\leq f_k(x|1,t)$. Moreover, for every $t=0,1,2,\dots$ there exists an integrable function $u(x|t)$, such that

$$f_k(x|s,t)\leq u(x|t),$$

for large values of k. For instance, the function $u(x|t)=10^t\,\mathrm{Ga}(x|0.1,1)$, where $\mathrm{Ga}(x|0.1,1)$ denotes the Gamma density with parameters 0.1 and 1, is an upper bound of $f_k(x|s,t)$.

Applying the dominated convergence theorem to the sequence $\{f_k(x|s,t), k\geq 1\}$, we have that

$$\lim_{k\to\infty}\int_0^1 f_k(x|s,t)\,dx = \int_0^1 \lim f_k(x|s,t)\,dx = 0,$$

and this completes the posterior model consistency proof for the Bayes factor based on the mixture of g-priors and the hierarchical uniform prior.

A similar proof can be given for the Bayes factors for $g=n$ and for the intrinsic priors. This completes the proof of Theorem 3.

ACKNOWLEDGMENTS

We are very grateful to two anonymous referees for very valuable comments that helped to improve the original manuscript. This paper has been supported by grant MTM2011-28945, Ministerio de Ciencia, Spain.

REFERENCES

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

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

BERGER, J. O. and PERICCHI, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (P. Lahiri, ed.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 135–207. IMS, Beachwood, OH. MR2000753

CASELLA, G. and MORENO, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.* **101** 157–167. MR2268035

CASELLA, G., GIRÓN, F. J., MARTÍNEZ, M. L. and MORENO, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37** 1207–1228. MR2509072

CLYDE, M. and GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 681–698. MR1796285

CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94. MR2082148

CLYDE, M., PARMIGIANI, G. and VIDAKOVIC, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85** 391–401. MR1649120

CONSONNI, G., FORSTER, J. J. and LA ROCCA, L. (2015). The whetstone and the alum block: Balanced objective Bayesian comparison of nested models for discrete data. *Statist. Sci.* **28** 398–423.

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

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

FRASER, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26** 299–316. MR2918001

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

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GIRÓN, F. J., MARTÍNEZ, M. L., MORENO, E. and TORRES, F. (2006). Objective testing procedures in linear models: Calibration of the p-values. *Scand. J. Stat.* **33** 765–784. MR2300915

GIRÓN, F. J., MORENO, E., CASELLA, G. and MARTÍNEZ, M. L. (2010). Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math. RACSAM* **104** 57–67. MR2666441

HANSEN, M. H. and YU, B. (2001). Model selection and the principle of minimum description length. *J. Amer. Statist. Assoc.* **96** 746–774. MR1939352

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074

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

LEON-NOVELO, L., MORENO, E. and CASELLA, G. (2012). Objective Bayes model selection in probit models. *Stat. Med.* **31** 353–365. MR2879809

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

MORENO, E. (1997). Bayes factors for intrinsic and fractional priors in nested models. Bayesian robustness. In L_1-*Statistical Procedures and Related Topics* (*Neuchâtel*, 1997). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **31** 257–270. IMS, Hayward, CA. MR1833592

MORENO, E., BERTOLINO, F. and RACUGNO, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *J. Amer. Statist. Assoc.* **93** 1451–1460. MR1666640

MORENO, E. and GIRÓN, F. J. (2005). Consistency of Bayes factors for intrinsic priors in normal linear models. *C. R. Math. Acad. Sci. Paris* **340** 911–914. MR2152278

MORENO, E. and GIRÓN, F. J. (2008). Comparison of Bayesian objective procedures for variable selection in linear regression. *TEST* **17** 472–490. MR2470092

MORENO, E., GIRÓN, F. J. and CASELLA, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Ann. Statist.* **38** 1937–1952. MR2676879

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

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682

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

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