The Annals of Statistics
2010, Vol. 38, No. 4, 1937–1952
DOI: 10.1214/09-AOS754

CONSISTENCY OF OBJECTIVE BAYES FACTORS AS THE MODEL DIMENSION GROWS

BY ELÍAS MORENO[1], F. JAVIER GIRÓN AND GEORGE CASELLA[2]

University of Granada, University of Málaga and University of Florida

In the class of normal regression models with a finite number of regressors, and for a wide class of prior distributions, a Bayesian model selection procedure based on the Bayes factor is consistent [Casella and Moreno *J. Amer. Statist. Assoc.* **104** (2009) 1261–1271]. However, in models where the number of parameters increases as the sample size increases, properties of the Bayes factor are not totally understood. Here we study consistency of the Bayes factors for nested normal linear models when the number of regressors increases with the sample size. We pay attention to two successful tools for model selection [Schwarz *Ann. Statist.* **6** (1978) 461–464] approximation to the Bayes factor, and the Bayes factor for intrinsic priors [Berger and Pericchi *J. Amer. Statist. Assoc.* **91** (1996) 109–122, Moreno, Bertolino and Racugno *J. Amer. Statist. Assoc.* **93** (1998) 1451–1460].

We find that the the Schwarz approximation and the Bayes factor for intrinsic priors are consistent when the rate of growth of the dimension of the bigger model is $O(n^b)$ for $b < 1$. When $b = 1$ the Schwarz approximation is always inconsistent under the alternative while the Bayes factor for intrinsic priors is consistent except for a small set of alternative models which is characterized.

1. Introduction. Statistical methodology based on Bayes factors is particularly suitable for dealing with multiple hypotheses testing problems when the dimension of the parameter spaces varies across models. In such cases, Bayesian and frequentist model selection procedures do not necessarily agree as the typically *ad hoc* dimension corrections of the different frequentist criteria do not provide the same results as those automatically produced by the Bayesian procedures which select models according to the parsimony principle. For a recent discussion on the topic see Girón et al. (2006).

In the class of normal linear regression models, consistency of Bayesian variable selection procedures, and, in particular, those using intrinsic priors, has been recently established in Casella et al. (2009). There it was shown that, under mild

Received March 2009; revised July 2009.

[1]Supported by Ministerio de Ciencia y Tecnología, Grant SEJ-65200 and Junta de Andalucía Grant SEJ-02814.

[2]Supported by NSF Grants DMS-04-05543, DMS-0631632 and SES-0631588.

AMS 2000 subject classifications. Primary 62F05; secondary 62J15.

Key words and phrases. Bayes factors, BIC, intrinsic priors, linear models, multiplicity of parameters, rate of growth.

regularity conditions, when sampling from a given submodel of a regression model with p regressors, the probability of selecting the true model tends to one as the sample size n tends to infinity, and the probability of selecting any other submodel tends to zero. It was also shown that the Schwarz (1978) approximation is, in spite of its simplicity, an accurate tool for selecting linear models when there is a small number of parameters and the sample size is moderate or large. Those results were obtained for a fixed number of regressors p, and hence a finite number of models. Other forms of consistency of Bayes factors for variable selection using Zellner's g-prior with several hyperpriors on g have been recently studied by Liang et al. (2008). These forms of consistency include the consistency when R^2 tends to one, when n tends to infinity, and consistency under prediction for squared error loss.

However, in some applications, the number of models increases with the sample size. For instance, clustering is an interesting model selection problem where the number of models increases as the sample size increases, and the question is whether consistency of the Bayesian model selection procedure based on intrinsic priors also holds in this latter context. Certainly, it will not be possible to consistently estimate the parameters of the underlying models, but we wonder whether consistently selecting the true model is still possible.

When the number of parameters increases with the sample size, an analysis of the consistency of several frequentist and Bayesian approximation criteria for model selection in linear models, including the Schwarz approximation, was given in Shao (1997). However, the results obtained by Shao (1997) do not coincide with ours, as the consistency notion used by Shao is not the same as the one we use here. Shao 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. We consider the true model to be the one from which the observations are drawn. Of course, the true model may not be in the class of models we are considering. In this case consistency does not hold although many Bayesian model selection procedures choose models in the class that are located as close as possible to the true one where closeness is related to a specific "natural" metric [Casella et al. (2009)].

We examine consistency in linear models of both the Bayes factors for intrinsic priors and the Schwarz approximation (BIC), when the dimension of the parameter space of the models increases with the sample size. We find that both the Bayes factor for the intrinsic priors and BIC are consistent under the null; however, they might be inconsistent under some alternative sampling models. The consistency depends on the rate of divergence of the dimension of the null and the alternative linear models. Roughly speaking, the BIC and the Bayes factor for intrinsic priors are consistent when the rate of growth of the dimension of the full model goes to infinity as $O(n^b)$ for any $b < 1$.

When $b = 1$, the BIC is always inconsistent under the alternative while for the Bayes factor for intrinsic priors there is an inconsistency region which is located in a small neighborhood of the null model. This neighborhood is characterized in

terms of a "distance" to the null sampling model. In particular, for the case of the oneway ANOVA, where $b = 1$, the Bayes factor for intrinsic priors is not consistent for all alternative models. This finding is apparently in contradiction with the results of Berger, Ghosh and Mukhopadhyay (2003) who find "*that suitable Bayes factors will be consistent*," and hence induces an apparent paradox. However, in Section 4 we are able to resolve the apparent contradiction and find that the consistency result in Berger, Ghosh and Mukhopadhyay (2003) is obtained by using a normal prior centered at the null with variance tending to zero, a situation that typically is not obtained by the intrinsic priors. We also observe that consistency is obtained for this problem for priors that degenerate to a point mass.

The rest of the paper is organized as follows. In Section 2 we characterize the consistency of the BIC and the Bayes factor for intrinsic priors for the usual linear regression model for $b < 1$, demonstrate the inconsistency of BIC for $b = 1$ and characterize the small inconsistency region for the Bayes for intrinsic priors for $b = 1$. Section 3 presents some models where the results of Section 2 apply, and Section 4 resolves the apparent paradox with the results of Berger, Ghosh and Mukhopadhyay (2003). Section 5 provides a short, concluding discussion, and there is an Appendix with some technical material.

2. Consistency in linear models. In this section we give, for normal linear regression models with parameters increasing with the sample size, conditions under which the Bayes factor for intrinsic priors asymptotically selects the correct model. The finding is that the Bayes factor may not be a consistent model selector for all parameter values, depending on the rate of divergence of the dimension of models. When there is an inconsistency region, it is characterized in terms of a "distance" from the alternative to the null model. We also show that the BIC model selector is inconsistent when sampling from the full model if $b = 1$.

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be a vector of independent responses, $\mathbf{X}_p$ a design matrix of dimension $n \times p$, where p is the number of explanatory variables, and let $\mathbf{X}_i$ denote a submatrix of $\mathbf{X}_p$ whose dimensions are $n \times i$. We compare the reduced sampling model $N(\mathbf{y}|\mathbf{X}_i\alpha_i, \sigma_i^2\mathbf{I}_n)$, and the full model $N(\mathbf{y}|\mathbf{X}_p\beta_p, \sigma_p^2\mathbf{I}_n)$, where the regression parameter vectors $\alpha_i = (\alpha_1, \ldots, \alpha_i)'$, $\beta_p = (\beta_1, \ldots, \beta_p)'$ and the variance errors σ_i^2, σ_p^2, are unknown. Note that the reduced model is nested in the full model. The comparison is based on the Bayes factor of model M_p versus model M_i, and we remark that it cannot be computed by using the reference prior, the usual objective priors, since they are improper and hence defined up to an arbitrary multiplicative constant. The so-called intrinsic priors that are given below solve this difficulty [Berger and Pericchi (1996), Moreno, Bertolino and Racugno (1998)]. These objective priors have proven to behave very well for multiple testing problems [Casella and Moreno (2006)].

To derive the Bayes factor for intrinsic priors we start with the improper reference priors $\pi^N(\alpha_i, \sigma_i) = c_i/\sigma_i$ and $\pi^N(\beta_p, \sigma_p) = c_p/\sigma_p$ where c_i and c_p are

arbitrary positive constants, so we consider the following Bayesian models:

$$M_i : \left\{ N(\mathbf{y}|\mathbf{X}_i\alpha_i, \sigma_i^2\mathbf{I}_n), \pi^N(\alpha_i, \sigma_i) = \frac{c_i}{\sigma_i} \right\}$$

and

$$M_p : \left\{ N(\mathbf{y}|\mathbf{X}_p\beta_p, \sigma_p^2\mathbf{I}_n), \pi^N(\beta_p, \sigma_p) = \frac{c_p}{\sigma_p} \right\}.$$

Standard calculations [Moreno, Bertolino and Racugno (1998), and Girón et al. (2006)] yield the following intrinsic prior for (β_p, σ_p), conditional on (α_i, σ_i):

$$\pi^I(\beta_p, \sigma_p|\alpha_i, \sigma_i) = \frac{2\sigma_i}{\pi(\sigma_i^2 + \sigma_p^2)} N_p\Big(\beta_p|\tilde{\alpha}_i, (\sigma_i^2 + \sigma_p^2)\mathbf{W}_p^{-1}\Big),$$

where $\tilde{\alpha}_i' = (\alpha_i', \mathbf{0}')$, $\mathbf{0}'$ being the null vector of $p - i$ components, and $\mathbf{W}_p^{-1} = \frac{n}{p+1}(\mathbf{X}_p'\mathbf{X}_p)^{-1}$. Then, using the priors $\{\pi^N(\alpha_i, \sigma_i), \pi^I(\beta_p, \sigma_p|\alpha_i, \sigma_i)\}$, the Bayes factor for comparing the model M_p and M_i is

$$B_{pi}(\mathbf{y}) = \frac{2}{\pi}(p+1)^{(p-i)/2} \int_0^{\pi/2} \frac{\sin^{p-i}\varphi (n + (p+1)\sin^2\varphi)^{(n-p)/2}}{(n\mathcal{B}_{ip} + (p+1)\sin^2\varphi)^{(n-i)/2}} d\varphi, \tag{1}$$

where

$$\mathcal{B}_{ip} = \frac{RSS_p}{RSS_i} = \frac{\mathbf{y}'(\mathbf{I}_n - \mathbf{H}_p)\mathbf{y}}{\mathbf{y}'(\mathbf{I}_n - \mathbf{H}_i)\mathbf{y}}$$

and $\mathbf{H}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$, $j = i, p$, is the hat matrix.

We first extend the definition of distance from M_p to M_i regression models given in Casella et al. (2009) to account for models for which the number of parameters and the sample size increase to infinity and define the "distance" from M_p to M_i for a given sample size n as

$$\delta_{pi} = \frac{1}{\sigma_p^2}\beta_p' \frac{\mathbf{X}_p'(\mathbf{I}_n - \mathbf{H}_i)\mathbf{X}_p}{n}\beta_p.$$

The asymptotic performance of the Schwarz approximation is given in Theorem 1, and that of the Bayes factor for the intrinsic priors is given in Theorems 2 and 3, and Corollary 4. The proofs of these results depend on Lemma 1.

In what follows $\lim_{n\to\infty}[M]\ Z_n$ will denote the limit in probability of the random sequence $\{Z_n; n \geq 1\}$ under the assumption that we are sampling from model M. This model M will have a fixed parameter sequence. Further, we will need to use the doubly noncentral beta distribution with parameters $\upsilon_1/2$, $\upsilon_2/2$ and noncentrality parameters λ_1, λ_2. One way to define this distribution is as follows. If Y_1, Y_2 are independent random variables with noncentral chi square distributions $\chi^2(y_1|\upsilon_1, \lambda_1)$ and $\chi^2(y_2|\upsilon_2, \lambda_2)$, respectively, then the variable $X = Y_1/(Y_1 + Y_2)$ follows the doubly noncentral beta distribution $\mathrm{Be}(\upsilon_1/2, \upsilon_2/2; \lambda_1, \lambda_2)$ [Johnson, Kotz and Balakrishnan (1995), page 502].

LEMMA 1.

1. *When sampling from model M_i the distribution of the statistics $\mathcal{B}_{ip}$ is the beta distribution* $\mathrm{Be}((n-p)/2, (p-i)/2)$, *and when sampling from model M_p it is the noncentral beta distribution* $\mathrm{Be}((n-p)/2, (p-i)/2); 0, n\delta_{pi})$.
2. *Let $\{X_n, n \geq 1\}$ be a sequence of random variables such that*

$$X_n \sim \mathrm{Be}\left(\frac{n-p}{2}, \frac{p-i}{2}; 0, n\delta_{pi}\right), \qquad n \geq 1.$$

If i and p vary with n as $i = O(n^a)$ and $p = O(n^b)$, where $0 \leq a \leq b \leq 1$, then:

(i) *If $a < b = 1$, when sampling from model M_p*

$$\lim_{n\to\infty} [M_p]X_n = \frac{1 - 1/r}{\delta + 1},$$

where the constant r satisfies $r = \lim_{p\to\infty} n/p > 1$, and $\delta = \lim_{n\to\infty} \delta_{pi}$.

(ii) *If $a = b = 1$, so there exist two positive constants such that $r = \lim_{p\to\infty} n/p > 1$ and $s = \lim_{p\to\infty} n/i > 1$, we have*

$$\lim_{n\to\infty} [M_p]X_n = \frac{1 - 1/r}{1 + \delta - 1/s}.$$

(iii) *If $b < 1$,*

$$\lim_{n\to\infty} [M_p]X_n = \frac{1}{1+\delta}.$$

PROOF. See the Appendix. □

2.1. *Inconsistency of BIC.* In this linear model setting we now prove that the Schwarz approximation for comparing M_p against M_i is inconsistent when sampling from M_p under certain conditions as first noticed by Stone (1979) in a special case.

THEOREM 1. *For comparing model M_p to model M_i, where M_i is nested in M_p, and $i = O(n^a)$ and $p = O(n^b)$, if $0 < a \leq b < 1$, the Schwarz approximation,*

$$S_{pi}(\mathbf{y}) = \exp\left\{\frac{i-p}{2}\log n - \frac{n}{2}\log \mathcal{B}_{ip}\right\}$$

is consistent under the null and the alternative. However, if $b = 1$ it is inconsistent under any alternative model M_p provided that $\lim_{n\to\infty} \delta_{pi} > 0$.

PROOF. Consistency under the null for both cases follows from part 1 of Lemma 1. For $b < 1$, we notice that the leading term of the Bayes factor is the

one involving the statistic $\mathcal{B}_{ip}(\mathbf{y})$, but from part (iii) of Lemma 1 the limit of the sequence $\mathcal{B}_{ip}(\mathbf{y})$ is a number strictly smaller than 1, and, therefore,

$$\lim_{n\to\infty}[M_p]S_{pi}(\mathbf{y}) = \infty.$$

On the other hand, if $b = 1$, then $p = n/r$ and $i = n/s$, where r is a positive number greater than 1 and s is a number greater than r, the leading term of the exponent of the Schwarz approximation is now the first one which is strictly negative. Therefore,

$$\lim_{n\to\infty}[M_p]S_{pi}(\mathbf{y}) = 0$$

and the proof is complete. □

2.2. *Consistency of the Bayes factors for intrinsic priors.* We now characterize the consistency of the Bayes factor for intrinsic priors, first assuming that both p and n increase at the same rate; that is, $r = \lim_{n\to\infty, p\to\infty} n/p$, is a strictly positive number. We further assume that the limit of the distance δ_{pi} is finite when p and n tend to infinity, and i is either finite or increases to infinity at a lower rate than n. Note that in this theorem the constant $b = 1$.

THEOREM 2. *Suppose that, as the sample size increases, models increase their number of parameters with rate* $i = O(n^a)$ *and* $p = O(n)$, *where* $0 \le a < 1$, *and* $r = \lim_{n,p\to\infty} n/p > 1$.

1. *When sampling from the simpler model* M_i, $\lim_{n\to\infty}[M_i]B_{pi}(\mathbf{y}) = 0$.
2. *When sampling from the alternative model* M_p *there exists a function* $\delta(r)$ *such that*

$$\lim_{n\to\infty}[M_p]B_{pi}(\mathbf{y}) = \begin{cases} \infty, & \text{if } \lim_{n\to\infty}\delta_{pi} > \delta(r), \\ 0, & \text{if } \lim_{n\to\infty}\delta_{pi} < \delta(r). \end{cases} \tag{2}$$

Further, this function has the simple expression

$$\delta(r) = \frac{r-1}{(r+1)^{(r-1)/r} - 1} - 1 \tag{3}$$

and is a decreasing convex function such that $\lim_{r\to\infty}\delta(r) = 0$.

PROOF. We first prove consistency of $B_{pi}(\mathbf{y})$ under the simpler model M_i. The Bayes factor B_{pi} in (1) can be written as

$$B_{pi}(\mathbf{y}) = \frac{2}{\pi}\int_0^{\pi/2}\left(1 + \frac{n}{(p+1)\sin^2\varphi}\right)^{(n-p)/2}\left(1 + \frac{n\mathcal{B}_{ip}}{(p+1)\sin^2\varphi}\right)^{-(n-i)/2} d\varphi.$$

From Lemma 1 it follows that

$$\lim_{p\to\infty}[M_i]\mathcal{B}_{ip} = \frac{r-1}{r}$$

and, replacing n by pr, the Bayes factor for large p can be approximated by

$$B_{pi}(\mathbf{y}) \approx \frac{2}{\pi}\int_0^{\pi/2}\left(1+\frac{r}{\sin^2\varphi}\right)^{p(r-1)/2}\left(1+\frac{r-1}{\sin^2\varphi}\right)^{(i-pr)/2}d\varphi.$$

As the integrand is a monotonic increasing function of the angle φ, the sup is attained at $\varphi=\pi/2$, and, therefore, an upper bound on the integrand is $(1+r)^{p(r-1)/2}r^{(i-pr)/2}$. Then, for large p, an upper bound for the Bayes factor is

$$B_{pi}(\mathbf{y}) < \left[\frac{(1+r)^{r-1}}{r^r}\right]^{p/2} r^{i/2}.$$

As the function of r enclosed in square brackets is strictly smaller than 1 for $r>1$, and the rate of growth of i is strictly smaller than that of p, it follows that

$$\lim_{p\to\infty}\left[\frac{(1+r)^{r-1}}{r^r}\right]^{p/2} r^{i/2}=0$$

for all $r>1$, thus proving consistency of the Bayes factor for the intrinsic prior under the reduced model M_i.

Consistency under the full model M_p is established as follows. From Lemma 1, the limiting distribution of the statistics $\mathcal{B}_{ip}$ under M_p is

$$\lim_{p\to\infty}[M_p]\mathcal{B}_{ip}=\frac{1-1/r}{\delta+1},$$

where δ is the limit of the the "distance" from the full model to the reduced one, which only depends on the limiting behavior of the parameters of the full model; that is,

$$\delta=\lim_{p\to\infty}\delta_{pi}=\lim_{p\to\infty}\frac{1}{\sigma_p^2}\beta_p'\frac{\mathbf{X}_p'(\mathbf{I}_n-\mathbf{H}_i)\mathbf{X}_p}{pr}\beta_p.$$

Therefore, the Bayes factor $B_{pi}(\mathbf{y})$ for large values of p can be approximated by

$$B_{pi}(\mathbf{y}) \approx \frac{2}{\pi}\int_0^{\pi/2}\left(1+\frac{r}{\sin^2\varphi}\right)^{p(r-1)/2}\left(1+\frac{r-1}{(1+\delta)\sin^2\varphi}\right)^{(i-pr)/2}d\varphi.$$

We look at two cases, depending on the values of the parameter δ.

For $\delta>1$, the Bayes factor is an increasing convex function of p and this implies that the Bayes factor is always consistent.

For $\delta\le 1$, the argument proceeds as follows. As the integrand is a continuous increasing function of φ for all r, δ and p, then by the mean value theorem, there exists a unique value of φ_0, say $0\le\varphi_0(r,p,\delta)\le\pi/2$, such that for large p the Bayes factor is approximated by

$$B_{pi}(\mathbf{y}) \approx \left(1+\frac{r}{\sin^2\varphi_0(r,p,\delta)}\right)^{p(r-1)/2}\left(1+\frac{r-1}{(1+\delta)\sin^2\varphi_0(r,p,\delta)}\right)^{(i-pr)/2}.$$

The limit of the sequence $\{\varphi_0(r, p, \delta), p \geq 1\}$ is seen to be equal to $\pi/2$ for all r, and $\delta \leq 1$, Thus, for large values of p, recalling that $i = o(p^b)$, we can further approximate the Bayes factor by

$$B_{pi}(\mathbf{y}) \approx \left[(1+r)^{r-1}\left(1+\frac{r-1}{1+\delta}\right)^{-r}\right]^{p/2}. \tag{4}$$

(It can be checked numerically that even for moderate values of p this approximation is very accurate.) Note that when the expression in square brackets is greater than 1 consistency holds, and when smaller than 1 the Bayes factor is inconsistent. The root of the equation

$$(1+r)^{r-1}\left(1+\frac{r-1}{1+\delta}\right)^{-r} = 1,$$

is $\delta(r)$ of (3), proving the theorem. □

We remark that the function $\delta(r)$ only depends on the lim $n/p = r$. In addition to the limiting value as $r \to \infty$, we also have $\lim_{r\to 0}\delta(r) = (e-1)^{-1}$ and $\lim_{r\to 1}\delta(r) = [log(2)]^{-1} - 1$. Notice that the case of equality in the limit (2) is not covered by the theorem. It happens that, in this case, we cannot make a specific conclusion as there will be parameter values for which there is, and there is not, consistency.

Theorem 2 covers the case in which the dimension of the parameter space grows at a rate strictly smaller than that of the sample space. However, it does not cover the case where the dimension of the null and the alternative space grow at the same rate as the sample size. This case is covered in Theorem 3.

THEOREM 3. *Suppose that, as the sample size increases, the rates the models increase their number of parameters are $i = O(n)$ and $p = O(n)$, and there exists positive constants r and s such that $r = \lim_{n,p\to\infty} n/p$ and $s = \lim_{n,i\to\infty} n/i \geq 1$.*

1. *When sampling from the simpler model M_i, $\lim_{n\to\infty}[M_i]B_{pi}(\mathbf{y}) = 0$.*
2. *When sampling from the alternative model M_p, there exists a function $\delta(r, s)$ such that*

$$\lim_{n\to\infty}[M_p]B_{pi}(\mathbf{y}) = \begin{cases} \infty, & \text{if } \lim_{n\to\infty}\delta_{pi} > \delta(r,s), \\ 0, & \text{if } \lim_{n\to\infty}\delta_{pi} < \delta(r,s). \end{cases}$$

This function has the following simple explicit form:

$$\delta(r,s) = \frac{r-1}{(r+1)^{s(r-1)/(r(s-1))} - 1} - 1 + \frac{1}{s} \tag{5}$$

and it is a bounded decreasing convex function in r for fixed s with $\delta(r,s) \leq 1/\log 2 - 1$ for all $s > r > 1$, and $\lim_{r\to\infty}\delta(r,s) = 0$ for all s. Further, $\lim_{s\to\infty}\delta(r,s) = \delta(r)$ of (3).

PROOF. To prove consistency under the simple model M_i, from Lemma 1 it follows that

$$\lim_{p\to\infty}[M_i]\mathcal{B}_{ip} = \frac{s(r-1)}{r(s-1)}$$

and, replacing n by pr, and $i/s = pr/s$, the Bayes factor for large p can be approximated by

$$B_{pi}(\mathbf{y}) \approx \frac{2}{\pi}\int_0^{\pi/2}\left(1+\frac{r}{\sin^2\varphi}\right)^{p(r-1)/2} \times\left(1+\frac{s(r-1)}{(s-1)\sin^2\varphi}\right)^{-pr(s-1)/(2s)} d\varphi. \tag{6}$$

As the integrand is a monotonic increasing function of the angle φ, the supremum is attained at $\varphi = \pi/2$, and thus an upper bound of the integrand is

$$(1+r)^{p(r-1)/2}\left(1+\frac{s(r-1)}{(s-1)}\right)^{-pr(s-1)/(2s)}.$$

Then, for large p, the Bayes factor is bounded from above by

$$B_{pi}(\mathbf{y}) < \left[(1+r)^{r-1}\left(\frac{rs-1}{s-1}\right)^{-r(s-1)/s}\right]^{p/2},$$

but as the function of r and s enclosed in square brackets is strictly smaller than 1 for $s > r > 1$, it follows that the limit of the upper bound of the Bayes factor is 0 for all $s > r > 1$ thus proving consistency of the Bayes factor for the intrinsic prior under the reduced model M_i.

Consistency under the full model M_p is proven in a similar way to that of Theorem 2. From Lemma 1, the limiting distribution of the statistics $\mathcal{B}_{ip}$ under M_p is now

$$\lim_{p\to\infty}[M_p]\mathcal{B}_{ip} = \frac{1-1/r}{1+\delta-1/s},$$

where δ is the same as in Theorem 3. Following the same course of reasoning as in Theorem 2, we finally arrive at the following new approximation for the Bayes factor for large values of p:

$$B_{pi}(\mathbf{y}) \approx \left[(1+r)^{r-1}\left(1+\frac{r-1}{1+\delta-1/s}\right)^{r(s-1)/s}\right]^{p/2}.$$

As the expression in square brackets does not depend on p, the limiting behavior of the Bayes factor depends on whether this expression is less than or greater than 1. Therefore, the new value of the boundary for consistency-inconsistency,

$\delta(r, s)$, is the root of the equation

$$(1+r)^{r-1}\left(1+\frac{r-1}{1+\delta-1/s}\right)^{r(s-1)/s}=1,$$

which is (5). This proves the theorem. □

REMARK 1. For all $s \geq r > 1$, the function $\delta(r, s)$ is bounded by a number smaller than 1. Note also that if the rate of growth of M_i is smaller than that of M_p, that is, $s \to \infty$, then it is easy to show that $\lim_{s\to\infty} \delta(r, s) = \delta(r)$.

An extension of Theorem 3 to the case where models M_i and M_p grow at a slower rate than the sample size; that is, $i = O(n^a)$ and $p = O(n^b)$, where $0 \leq a = b < 1$, can be regarded as a limiting case of the preceding theorem where both r and s go to infinity. So, we have the following corollary.

COROLLARY 4. *For $i = O(n^a)$ and $p = O(n^b)$ and $0 \leq a \leq b < 1$, the Bayes factor for intrinsic priors is consistent if* $\lim_{n\to\infty} \delta_{pi} > 0$.

3. Applications. We look at some practical models for which the results of the preceding section can be applied, including various ANOVA models, the multiple change point problem, the clustering problem and spline regression.

In particular, the classical ANOVA problem will be illustrated in some detail. For instance, we will see that for the one-way ANOVA, and by extension any full factorial completely randomized design, the Bayes factor for intrinsic priors is inconsistent in a region around the null. However, reducing the ANOVA model by eliminating interaction terms recovers consistency.

3.1. *Homoscedastic ANOVA.* There is a subtle difference between an ANOVA with a full model specification (including all interactions) and one with a reduced model specification, as it results in different asymptotic rates. We present the results for balanced models with the same number of observations per cell, but they can easily be extended to cover the unbalanced case.

3.1.1. *Full model specification.* We give a detailed development for the one-way ANOVA, and then show how the results apply to full factorial designs. The null sampling model of the homoscedastic one-way ANOVA, M_1, where it is assumed that the means are equal to an unknown μ, can be written as

$$M_1 : \left\{ N(\mathbf{y}|\mu\mathbf{1}_n, \tau^2\mathbf{I}_n), \pi^N(\mu, \tau) = \frac{c}{\tau} \right\}$$

and the alternative model as

$$M_p : \left\{ N(\mathbf{y}|\mathbf{X}_p\mu_p, \sigma^2\mathbf{I}_n), \pi^I(\mu_p, \sigma|\mu, \tau) \right.$$

$$\left. = HC^+(\sigma|\mu, \tau)\prod_{i=1}^{p} N\left(\mu_i|\mu, \frac{\tau^2+\sigma^2}{2}\right) \right\},$$

where c is an arbitrary positive constant, $\mathbf{1}_n$ denotes a vector of n components containing 1's, $\mathbf{X}_p$ is an $n \times p$ matrix such that the first r rows are equal to the unit vector $\mathbf{e}_1$, the next r rows are equal to the unit vector $\mathbf{e}_2$ and so on, so that the last r rows are equal to the unit vector $\mathbf{e}_p$ where the unit vector $\mathbf{e}_j$ has coordinate 1 at the jth position, and $HC^+(\sigma|\mu,\tau)$ represents the half Cauchy prior density of σ, conditional on μ, τ, on the positive part of the real line.

Since the dimension of M_1 is 2 and the dimension of M_p is $n+1$, Theorem 2 shows that there is an inconsistency region given by those alternative models with $\lim_{p\to\infty} \delta_{p1} < \delta(r)$ where $\delta(r)$ is given in (3). Thus, when sampling from M_p we have that

$$\lim_{n\to\infty}[M_p]B_{p1}(\mathbf{y}) = \begin{cases} \infty, & \text{if } \lim\limits_{p\to\infty} \delta_{p1} > \delta(r), \\ 0, & \text{if } \lim\limits_{p\to\infty} \delta_{p1} < \delta(r), \end{cases}$$

where the distance δ_{p1} is given by

$$\delta_{p1} = \frac{1}{n\sigma^2}\mu_p\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n\right)\mu_p = \frac{1}{\sigma^2}\frac{1}{p}\sum_{i=1}^{p}(\mu_i - \bar{\mu}_p)^2.$$

If we have a multiway completely randomized design the same results hold, as such a design is equivalent to a one-way design. For example, suppose we have a three-way full factorial with the model

$$y_{ijk} = \mu_i + \tau_j + \gamma_k + (\mu\tau)_{ij} + (\mu\gamma)_{ik} + (\tau\gamma)_{jk} + (\mu\tau\gamma)_{ikj} + \varepsilon_{ijk}, \quad i = 1,\dots,I,\ j = 1,\dots,J,\ k = 1,\dots,K. \tag{7}$$

The number of parameters (with identifiability restrictions) is IJK, and thus we are again in the case of Theorem 2 with $b = 1$. Any null hypothesis will result in a model M_1 with a reduced set of parameters that will satisfy $a < b$ of Theorem 2. Thus when sampling from the full model, the intrinsic Bayes procedure is consistent only if $\delta_{p1} > \delta(r)$, where $\delta(r)$ is given in (3), and, analogous to the one-way case, δ_{p1} is equal to the sum of squares of the differences between the null model coefficients and the full model coefficients. Extension to higher-order designs is straightforward.

3.1.2. *Reduced model specification.* In higher-order ANOVA models, it is often the case that some interaction terms are not specified. In particular, if the highest order interaction is not in the model, we can attain consistency of the intrinsic Bayes factor over the entire parameter space. We illustrate this with the three-way model (7); the extension to higher-order models should be clear.

If we eliminate the term $(\mu\tau\gamma)_{ikj}$ from the model (7), then there are at most

$$p = I + J + K + IJ + IK + JK$$

parameters in the full model M_2. Since there are $n = rIJKL$ observations, it immediately follows that

$$p = \begin{cases} O(n), & \text{if } I \text{ or } J \text{ or } K \to \infty, \\ o(n), & \text{if } I \text{ and } J \text{ and } K \to \infty. \end{cases}$$

So in the first case we can apply Theorem 2, and, similar to the full model evaluation, there will be an inconsistency region. However, in the second case, when all of I, J, and $K \to \infty$ we are in the case of Corollary 4; there is no inconsistency region and the Bayes factor for the intrinsic priors is consistent in the entire parameter space.

3.2. *Nested regression models*. Clustering, multiple change points and spline regression are examples of model selection problems for which the dimension of the alternative models grows at the same rate as the sample size n. Therefore, in the notation of the preceding sections $b = 1$, and hence the Schwarz approximation is inconsistent, but the Bayes factor for intrinsic priors is consistent except for a small region around the null model. Note that the null model in clustering is the one cluster model, in the multiple change points problem the null model is the no change model, and in spline regression the null model is the model that specifies no knots.

4. Comparison with previous findings. As we have seen in Section 3.1.1, in the homoscedastic ANOVA there is a region of inconsistency for the Bayes factor for intrinsic priors. This result seems to be in contradiction with the finding by Berger, Ghosh and Mukhopadhyay (2003) who consider the Bayes factor for normal priors. The models they compare are essentially

$$\begin{aligned} M_1 &: \left\{ \prod_{i=1}^{p} \prod_{j=1}^{r} N(y_{ij}|0,1) \right\}, \\ M_2 &: \left\{ \prod_{i=1}^{p} \prod_{j=1}^{r} N(y_{ij}|\mu_i,1), \pi^I(\mu_p|t) = \prod_{i=1}^{p} N(\mu_i|0,2/t), t \geq 1 \right\}, \end{aligned} \tag{8}$$

where $\pi^I(\mu_p|t)$ is the intrinsic prior when a training sample of size t is considered. Observe that the hyperparameter t controls the degree of concentration of the intrinsic priors around the null, and it usually ranges from 1 to r so as to not exceed the concentration of the likelihood of μ_i [for a discussion on the topic see Casella and Moreno (2009)].

For a given sample $\mathbf{y} = \{y_{ij},\ j = 1, \ldots, r,\ i = 1, \ldots, p\}$, the Bayes factor for intrinsic priors to compare the Bayesian model M_2 against M_1 is

$$B_{21}(\mathbf{y}|t) = \left(\frac{t}{2r+t} \right)^{p/2} \exp\left\{ \frac{r^2}{2r+t} \sum_{i=1}^{p} \bar{y}_i^2 \right\}$$

and it satisfies

$$\lim_{p\to\infty}[M_2]B_{21}(\mathbf{y}|t)=\begin{cases}\infty, & \text{if } \lim_{p\to\infty}\dfrac{1}{p}\displaystyle\sum_{i=1}^{p}\mu_i^2 > R(t,r),\\ 0, & \text{if } \lim_{p\to\infty}\dfrac{1}{p}\displaystyle\sum_{i=1}^{p}\mu_i^2 < R(t,r),\end{cases} \tag{9}$$

where $R(t,r)=(2r+t)(2r^2)^{-1}\ln[(2r+t)t^{-1}]-r^{-1}$, $1\le t\le r$.

As a curiosity we mention that the function $R(t,r)$ is related to the function $\delta(r)$ of Theorem 2 in the following way:

$$R(2,r)<\delta(r)<R(1,r),\qquad r\ge 1.$$

The Bayes factor for intrinsic priors is not consistent for all possible alternative sampling models, and thus we cannot call it a consistent model selector. For each t and r, the inconsistency region in the alternative parametric space will be denoted as

$$C(t,r)=\left\{\mu:0<\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\mu_i^2<R(t,r)\right\}. \tag{10}$$

We note that the bound $R(t,r)$ is a decreasing function in both arguments t and r, and $\lim_{r\to\infty}R(t,r)=\lim_{t\to\infty}R(t,r)=0$.

It turns out that for some extreme priors the Bayes factor is a consistent model selector. We present two extreme cases: the first one where the prior degenerates to a point mass, and the second one for intrinsic priors with variances that tend to zero.

1. *Simple null versus simple alternative*. As a modification of (8), suppose we want to choose between

$$M_1:\prod_{i=1}^{p}\prod_{j=1}^{r}N(y_{ij}|0,1)\quad\text{and}\quad M_2:\prod_{i=1}^{p}\prod_{j=1}^{r}N(y_{ij}|\mu_{i0},1),$$

where $\{\mu_{i0}, i\ge 1\}$ is an arbitrary but specified sequence such that $\lim_{p\to\infty}\sum\mu_{i0}^2/p>0$. Then, the Bayes factor $B_{21}(\mathbf{y}_p)$ satisfies $\lim_{p\to\infty}[M_2]B_{21}(\mathbf{y})=\infty$; that is, the Bayes factor is consistent under the alternative. This simple result means that when the prior distribution on the alternative degenerates to a point mass, consistency of the corresponding Bayes factor holds.

2. *Mixture priors*. The presence of uncertainty in the alternative models provokes the appearance of an inconsistency region $C(t,r)$. However, in Berger, Ghosh and Mukhopadhyay (2003), they use a continuous version of the intrinsic prior above and augment M_2 by mixing the variance $1/t$ of the $N(\mu_i|0,1/t)$ with a hyperprior density $g(t)$. (Special cases they consider are to take g to be either gamma or beta, yielding priors that they refer to as Cauchy and Smooth Cauchy.) For these general mixture priors they prove the following theorem.

THEOREM [Berger, Ghosh and Mukhopadhyay (2003), Theorem 3.1]. *For any prior of the form*

$$\pi_g(\mu) = \int_0^\infty \frac{t^{p/2}}{(2\pi)^{p/2}} e^{-(t/2)\sum_i \mu_i^2} g(t)\, dt \tag{11}$$

with $g(t)$ having support on $(0, \infty)$, the Bayes factor is consistent under M_1. Consistency under M_2 holds if

$$\tau^2 = \lim_{p\to\infty} \frac{1}{p} \sum_i \mu_i^2 > 0. \tag{12}$$

How do we reconcile (9) and (12), an apparent paradox? To obtain consistency for any alternative sampling model, we need the function in (9) to be zero, but this only occurs when t goes to infinity because r is fixed. Since the inconsistency regions $\{C(t, r), t \geq 1\}$ form a monotone decreasing sequence, the limit is $C_\infty(r) = \bigcap_{t=1}^\infty C(t, r) = \{\mu_i : \lim_{p\to\infty} \frac{1}{p} \sum_i \mu_i^2 = 0\}$, a point that does not belong to the alternative parameter space. In the above theorem this is exactly what the prior $\pi_g(\mu)$ does by incorporating priors with variance that tends to zero. (Something similar produces the so-called Lindley paradox when testing that the mean of a normal is zero; as the variance of the normal prior goes to zero less and less prior mass is given to any neighborhood of the null.)

Certainly, if we mix values of t from 1 to $r < \infty$, for instance mixing all the intrinsic priors, the intersection of these inconsistency regions $C_r(r) = \bigcap_{t=1}^r C(t, r)$, is a nonempty set in the alternative model space, and hence the inconsistency region does not disappear. This is also noted by Berger, Ghosh and Mukhopadhyay in their Theorem 3.2. that we state here using our notation.

THEOREM [Berger, Ghosh and Mukhopadhyay (2003), Theorem 3.2]. *For any prior of form* (11), *with $g(t)$ being supported on a finite interval* $[0, 1]$, *and* $r = 1$, *the Bayes Factor is inconsistent under M_2 for* $0 < \tau^2 < 2\log 2 - 1$.

We note that here $t = 2$ and $R(2, 1) = 2\log 2 - 1$.

5. Discussion. In our previous work [Casella et al. (2009)], where we looked at consistency of Bayes factors for a fixed number of parameters, we found that both the Bayes factor for intrinsic priors and the Schwarz approximation to a Bayes factor had the same asymptotic behavior, and both were consistent. In this paper we have derived the asymptotic behavior of the Bayes factor for intrinsic priors and the Schwarz approximation when the dimension of the model grows with the sample size, and we note an interesting dichotomy in their performance. The Bayes factor for intrinsic priors and the Schwarz approximation have very different asymptotic behavior for the usual case where the dimension of the full model grows at the

TABLE 1

Rate of divergence	Consistency region of $B_{pi}(\mathbf{y})$
$0 < a = b = 1$	$M_p : \lim_{n\to\infty} \delta_{pi} > \delta(r, s)$
$0 \le a < b = 1$	$M_p : \lim_{n\to\infty} \delta_{pi} > \delta(r)$
$0 \le a \le b < 1$	$M_p : \lim_{n\to\infty} \delta_{pi} > 0$

same rate as the sample size with the Bayes factor for intrinsic priors clearly being the optimal one.

We summarize the consistency regions of the Bayes factor for intrinsic priors for different values of a and b in Table 1, and we extract the following recommendations. For models with $b < 1$, the existence of very many parameters is not an inconvenience as far as consistency is concerned. For models with $b = 1$, there is a small inconsistency region around the null defined by the function δ that decreases rapidly as r increases. It also follows that inconsistency is the exception for the Bayes factor for intrinsic priors, while the rule is consistency, and this gives credence to the Bayes factor for intrinsic priors as a powerful objective tool for model selection.

APPENDIX: PROOF OF LEMMA 1

Part 1 follows from Theorem 1 in Casella et al. (2009). To prove part 2, we note that X_n can be written as

$$X_n = \left(1 + \frac{V_n}{W_n}\right)^{-1},$$

where $V_n \sim (1/n)\chi^2_{p-i}(n\delta_{pi})$ and $W_n \sim (1/n)\chi^2_{n-p}$. The means and variances of these random variables are

$$\mathrm{E}(V_n) = \delta_{pi} + \frac{p-i}{n}, \qquad \mathrm{E}(W_n) = 1 - \frac{p}{n}$$

and

$$\mathrm{Var}(V_n) = \frac{4\delta_{pi}}{n} + \frac{2(p-i)}{n^2}, \qquad \mathrm{Var}(W_n) = \frac{2(n-p)}{n^2}.$$

From these expressions the three cases follow:

(i) If $a < b = 1$, then when sampling from model M_p

$$V_n \to \delta + \frac{1}{r}, \qquad W_n \to 1 - \frac{1}{r} \quad \text{and} \quad X_n \to \frac{1 - 1/r}{1 + \delta}.$$

(ii) If $a = b = 1$, then when sampling from model M_p

$$V_n \to \delta + \frac{1}{r} - \frac{1}{s}, \qquad W_n \to 1 - \frac{1}{r} \quad \text{and} \quad X_n \to \frac{1 - 1/r}{1 + \delta - 1/s}.$$

(iii) If $b < 1$, then when sampling from model M_p

$$V_n \to \delta, \qquad W_n \to 1 \quad \text{and} \quad X_n \to \frac{1}{1+\delta}.$$

REFERENCES

BERGER, J. O., GHOSH, J. K. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112** 241–258. MR1961733

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

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

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

CASELLA, G. and MORENO E. (2009). Assessing robustness of intrinsic test of independence in two-way contingency tables. *J. Amer. Statist. Assoc.* **104** 1261–1271.

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. Statist.* **33** 765–784. MR2300915

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1995). *Continuous Univariate Distributions*, 2nd ed. Wiley, New York. MR1326603

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

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

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

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

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc. Ser. B* **41** 276–278.

E. MORENO
DEPARTMENT OF STATISTICS
UNIVERSITY OF GRANADA
18071, GRANADA
SPAIN
E-MAIL: emoreno@ugr.es

F. J. GIRÓN
DEPARTMENT OF STATISTICS
UNIVERSITY OF MÁLAGA
MALAGA
SPAIN
E-MAIL: giron@uma.es

G. CASELLA
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: casella@stat.ufl.edu