# On consistency of Bayesian variable selection procedures

Elias Moreno, Javier Girón, George Casella, Lina Martínez, F. J. Vázquez–Polo and María Martel

**Abstract** In this paper we extend the pairwise consistency of the Bayesian procedure to the entire class of linear models when the number of regressors grows as the sample size grows, and it is seen that for establishing consistency both the prior over the model parameters and the prior over the models play now an important role. We will show that commonly used Bayesian procedures with non–fully Bayes priors for models and for model parameters are inconsistent, and that fully Bayes versions of these priors correct this undesirable behavior.

**Key words:** Bayes factor, Bayesian information criterion (BIC), consistency in the class of linear models, $g-$priors, intrinsic priors, rate of growth of the number of regressors, variable selection.

## 1 Introduction

In some applications of the regression models the number of regressors grows as the sample size grows; for instance, clustering is an interesting model selection problem where the number of models increases as the sample size increases. The question

E. Moreno
Department of Statistics and Operation Research, University of Granada, Spain, e-mail: emoreno@ugr.es

F.J. Girón and M.L. Martínez
Department of Statistics and Operation Research, University of Málga, Spain, e-mail: {fj_giron or mlmartinez}@uma.es

G. Casella
Department of Statistics, University of Florida, USA, e-mail: casella@stat.ufl.edu

F.J. Vázquez–Polo and M. Martel
Department of Quantitative Methods, University of Las Palmas de Gran Canaria, Spain, e-mail: {fjvpolo or mmartel}@dmc.ulpgc.es

is whether consistency of the Bayesian variable selection procedure holds in this context. A partial answer to this question was given in Moreno et al. (2010), where consistency of the Bayes factor (pairwise consistency) for nested models when the number of regressor $k$ 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 a consistent procedure 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 asymptotic behavior.

Nevertheless, variable selection in regression is carried out in the entire class of normal regression models $\mathfrak{M}$ that contains nested and also nonnested models, and we wonder if the pairwise consistency when $k = O(n^b)$, $b \leq 1$, can be extended to the class $\mathfrak{M}$. We shall show here that the answer to this question depends not only on the Bayes factor but also on the prior over the class of models $\mathfrak{M}$. In fact, some commonly used Bayes factors and priors over models in the case of finite $k$, provide inconsistent Bayesian variable selection procedures in the class $\mathfrak{M}$ as $k$ grows with $n$.

## 2 Background

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

$$Y = \alpha_0 + \alpha_1 X_1 + ... + \alpha_k X_k + \varepsilon_k, \ \ \varepsilon_k \sim N(0, \sigma_k^2),$$

where the vector of regression coefficients $\alpha_{k+1} = (\alpha_0, \alpha_1, ..., \alpha_k)'$ and the variance error $\sigma_k^2$ are unknown. For a dataset $(\mathbf{y}, \mathbf{X})$, we denote the full model as $M_k$ with sampling normal distribution $N_n(\mathbf{y}|\mathbf{X}_{k+1}\alpha_{k+1}, \sigma_k^2 \mathbf{I}_n)$, where $\mathbf{y}$ is a vector of dimension $n$ of independent observations of $Y$ and $\mathbf{X}_{k+1}$ a $n \times (k+1)$ design matrix of full rank that involves $k$ regressors. The intercept only model $N_n(\mathbf{y}|\mu_0 \mathbf{1}_n, \sigma_0^2 \mathbf{I}_n)$ will be denoted as $M_0$.

The class of regression models defined by all possible subsets of regressors of $\{X_1, ..., X_k\}$ will be denoted as $\mathfrak{M}$ the number of which is $2^k$. A generic sampling model in the class containing $j$ of the potential $k$ regressors with sampling density $N_n(\mathbf{y}|\mathbf{X}_{j+1}\beta_{j+1}, \sigma_j^2 \mathbf{I}_n)$, where $\beta_{j+1} = (\beta_0, \beta_1, ..., \beta_j)$ is an unknown vector of regression coefficients, $\mathbf{X}_{j+1}$ is the $n \times (j+1)$ design submatrix of $\mathbf{X}_{k+1}$ and $\sigma_j^2$ is the unknown variance error, will be denoted as $M_j$. There are $\binom{k}{j}$ such a $M_j$ models and this subclass is denoted as $\mathfrak{M}_j$. It is clear that $\mathfrak{M} = \cup_{j=0}^k \mathfrak{M}_j$. The developments in the paper will be clear using this somewhat ambiguous, but simpler, notation.

Given a dataset $(\mathbf{y}, \mathbf{X})$ coming from an unknown model $M_T$ in $\mathfrak{M}$, and the priors for models and model parameters $\{\pi(\beta_j, \sigma_j | M_j)\pi(M_j), M_j \in \mathfrak{M}\}$, the model posterior probability of model $M_j$, which is used as the variable selector, is formally given by

$$\Pr(M_j|\mathbf{y},\mathbf{X}) = \frac{B_{j0}(\mathbf{y},\mathbf{X})\ \pi(M_j)/\pi(M_0)}{1+\sum_{\substack{M_\gamma\in\mathfrak{M}\\ M_\gamma\neq M_0}} B_{\gamma 0}(\mathbf{y},\mathbf{X})\ \pi(M_\gamma)/\pi(M_0)},\ M_j\in\mathfrak{M},$$

where the Bayes factor $B_{j0}(\mathbf{y},\mathbf{X})$ is

$$B_{j0}(\mathbf{y},\mathbf{X}) = \frac{\int N_n(\mathbf{y}|\mathbf{X}_j\beta_j,\sigma_j^2\mathbf{I}_n)\pi(\beta_j,\sigma_j)d\beta_j d\sigma_j}{\int N_n(\mathbf{y}|\alpha_0,\sigma_0^2\mathbf{I}_n)\pi(\alpha_0,\sigma_0)d\alpha_0 d\sigma_0}.$$

The advantage of the above expression for the model posterior probability is that all the Bayes factors in it involves nested models. This approach is called encompassing from below variable selection (Girón et al., 2006). We coud also use the encompassing from above approach in which all the Bayes factors are of the form $B_{jk}(\mathbf{y},\mathbf{X})$ (Casella and Moreno 2006).

For the dataset $(\mathbf{y},\mathbf{X})$ the expressions of the Bayes factors for comparing $M_0$ versus $M_j$ for the $g-$priors with $g=n$, $B_{j0}$, for the mixture of $g-$priors with mixing distribution an inverse Gamma$(g|1/2,n/2)$, $B_{j0}^{Mix}$, and for the intrinsic prior, $B_{j0}^{IP}$, are given by

$$B_{j0} = \frac{(1+n)^{(n-j-1)/2}}{(1+n\mathscr{B}_{j0})^{(n-1)/2}}, \tag{1}$$

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

where the integral on $\mathbb{R}^+$ does not have an explicit expression and needs numerical integration on $(0,\infty)$, and

$$B_{j0}^{IP} = \frac{2}{\pi}(j+2)^{j/2}\int_0^{\pi/2} \frac{\sin^j\varphi\ (n+(j+2)\sin^2\varphi)^{(n-j-1)/2}}{(n\mathscr{B}_{j0}+(j+2)\sin^2\varphi)^{(n-1)/2}}\,d\varphi, \tag{3}$$

where the integral does not have an explicit expression and needs numerical integration on $(0,\pi/2)$. These three Bayes factors depend on the data through the same statistic $\mathscr{B}_{j0}$, which is the ratio of the square sum of the residuals of model $M_j$ and $M_0$, that is

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

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

The prior over the class of models we consider are known as the independent Bernoulli prior and are given by

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

where $\theta$ is an unknown hyperparameter the meaning of which is the probability that a regressor is included in the model. The rationale for this prior is that all the regressors have *a priori* the same probability of inclusion. Note that when $\theta = 1/2$ the uniform prior is obtained. If we assume a uniform distribution for $\theta$, and the class $\mathfrak{M}$ is decomposed as $\mathfrak{M} = \cup_{j=0}^{k} \mathfrak{M}_j$, a hierarchical uniform prior for the models is obtained, that is

$$\begin{aligned}\pi^{HU}(M_j) &= \pi^{HU}(M_j|\mathfrak{M}_j)\pi^{HU}(\mathfrak{M}_j) \\ &= \int_0^1 \theta^j (1-\theta)^{k-j} d\theta = \binom{k}{j}^{-1} \frac{1}{k+1}.\end{aligned}$$

This means that conditional on the class $\mathfrak{M}_j$ the model prior $\pi^{HU}(M_j|\mathfrak{M}_j)$ is uniform, and the marginal $\pi^{HU}(\mathfrak{M}_j)$ of the classes $\{\mathfrak{M}_j,\ j = 0, 1, ..., k\}$ is also uniform.

## 3 Consistency in the class $\mathfrak{M}$ when $k = O(n^b)$ for $b \leq 1$

Let us define the random variable $X_k = j/k,\ j = 0, ..., k$, that takes values in [0,1] and indicates the proportion of regressors with respect to $k$ in the class $\mathfrak{M}_j,\ j = 0, ..., k$, and having the probability distribution given by

$$\Pr[X_k = \frac{j}{k}] = \Pr(\mathfrak{M}_j|k, \mathbf{y}, \mathbf{X}),\ j = 0, ..., k.$$

**Definition 1.** A Bayesian procedure is star consistent when sampling from the null $M_0 \equiv \mathfrak{M}_0$, if

$$\lim_{n\to\infty} \Pr[X_k \leq \varepsilon] = 1,\ [M_0].$$

**Theorem 1.** *If we sample from $M_0$ and the rate of growing of the number of regressors $k$ is $k = 0(n^b)$ for any $b < 1$, we have that*

*(i) the Bayesian variable selection procedures given by the Bayes factors $B_{j0}$, $B_{j0}^{Mix}$ and $B_{j0}^{IP}$ and the hierarchical uniform prior over models $\{\pi^{HU}(M_j),\ M_j \in \mathfrak{M}\}$, are consistent in the class $\mathfrak{M}$.*

*(ii) Further, if the prior over models is the independent Bernoulli $\pi(M_j|\theta)$, $M_j \in \mathfrak{M}$, the Bayesian procedure for $B_{j0}$, is consistent for $b < 1/2$, inconsistent for $b = 1/2$, and star consistent for $b > 1/2$. The Bayesian procedures for $B_{j0}^{Mix}$ and $B_{j0}^{IP}$ are only star consistent*

**Theorem 2.** *Assuming that* $\lim_{n\to\infty} n/j = r > 1$, *that is $k = 0(n)$, we have that:*

*(i) The Bayes factor $B_{j0}$ satisfies*

$$\lim_{n\to\infty} B_{j0} = \begin{cases} 0,\ [M_0], \\ 0,\ [M_j]. \end{cases}$$

*(ii) The Bayes factor $B_{j0}^{Mix}$ is such that*

$$\lim_{n\to\infty} B_{j0}^{Mix} = \begin{cases} 0, \ [M_0], \\ 0, \ \mathit{if} \ \ \delta_{j0} < \delta_{Mix}(r), \ [M_j], \\ \infty, \mathit{if} \ \ \delta_{j0} > \delta_{Mix}(r), \ [M_j], \end{cases}$$

*where*

$$\delta_{Mix}(r) = \left(1 - \frac{1}{r}\right)(e\ r)^{1/(r-1)} - 1.$$

*(iii) The Bayes factor $B_{j0}^{IP}$ satisfies*

$$\lim_{n\to\infty} B_{j1}^{IP} = \begin{cases} 0, \ [M_0] \\ 0, \ \mathit{if} \ \ \delta_{j0} < \delta_{IP}(r), [M_j], \\ \infty, \mathit{if} \ \ \delta_{j0} > \delta_{IP}(r), \ [M_j], \end{cases}$$

*where*

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

Part (i) of this theorem means that when $j = O(n)$ the Bayes factor for the $g-$prior with $g = n$ asymptotically always chooses the null model, regardless the model from which we are sampling. Therefore, in the rest of this section we rule it out.

Parts (ii) and (iii) of this theorem means that for both Bayes factors $B_{j0}^{Mix}$ and $B_{j0}^{IP}$ there are small regions of alternative models around the null for which inconsistency holds. However, we can show that the inconsistency region of the former contains the inconsistency region of the latter so that $B_{j0}^{IP}$ improves the asymptotic behavior of $B_{j0}^{Mix}$, the more so for values of $r$ near 1.

**Lemma 1.** *The following properties hold:*

*(i) The inconsistency region of the Bayes factor for the intrinsic priors $B_{j0}^{IP}$ is strictly contained in the inconsistency region of the Bayes factor for the mixture of $g-$priors $B_{j0}^{Mix}$.*

*(ii) In particular, for $r = 1$ the Bayes factor $B_{j0}^{Mix}$ is inconsistent under any alternative model, while the $B_{j0}^{IP}$ is consistent for $\delta_{j0} > 1/\log 2 - 1$.*

**Theorem 3.** *Assuming that the potential number of regressors k satisfies that* $\lim_{n\to\infty} n/k = s > 1$, *and the hierarchical uniform prior for models*

$$\pi^{HU}(M_j) = \frac{1}{k+1}\binom{k}{j}^{-1}, \ M_j \in \mathfrak{M},$$

*is used, the Bayesian variable selection procedures for either the Bayes factors $B_{j1}^{Mix}$ or $B_{j1}^{IP}$, $M_j \in \mathfrak{M}$, are consistent when sampling from $M_0$.*

# 4 Concluding remarks

When the potential number of regressors $k$ grows at a rate $O(n^b)$ for $0 \leq b \leq 1$, we have extended the pairwise consistency of Bayesian variable selection procedures to the consistency in the entire class of regression models $\mathfrak{M} = \cup_{j=1}^{k} \mathfrak{M}_j$. In this setting, the consistency of the posterior model probabilities depends not only on the Bayes factor but also on the prior over the class of models $\mathfrak{M}$.

The analyses have been carried out for three popular Bayes factors for model selection, which are based on the $g-$prior for $g = n$, on a mixture of $g-$priors and on the intrinsic priors; and the priors over the class of models: the well–known parametric independent Bernoulli $\{\pi^{IB}(M|\theta), 0 < \theta < 1\}$, and a mixture of them, the hierarchical uniform prior $\pi^{HU}(M)$, $M \in \mathfrak{M}$.

Different conclusions are drawn when the rate of growth of $k$ is $k = O(n^b)$ with $b < 1$, and when the rate is $k = O(n)$. In the former setting the conditions to ensure the consistency of the Bayesian procedure are less stringent than in the later setting. A summary of the conclusions on the consistency of the Bayesian procedures when sampling from $M_0$ and $k = O(n^b)$ is given in Table 1.

| $k = O(n^b)$ | $B_{j0}$ | | $B_{j0}^{Mix}$ and $B_{j0}^{IP}$ | |
|---|---|---|---|---|
| | $\pi(M\|\theta)$ | $\pi^{HU}(M)$ | $\pi(M\|\theta)$ | $\pi^{HU}(M)$ |
| $b < 1/2$ | Consistent | Consistent | Star Consistent | Consistent |
| $b = 1/2$ | Inconsistent | Consistent | Star Consistent | Consistent |
| $1/2 < b < 1$ | Star Consistent | Consistent | Star Consistent | Consistent |
| $b = 1$ | does not apply | does not apply | Inconsistent | Consistent |

**Table 1** Consistency of the Bayesian procedures when sampling from $M_0$ as a function of the Bayes factors and model prior.

A first conclusion we draw from Table 1 is that when sampling from $M_0$ and the prior over models is $\pi^{HU}(M)$, a fully–Bayes hierarchical uniform prior, the Bayesian procedures for the Bayes factors $B_{j0}, B_{j0}^{Mix}$, and $B_{j0}^{IP}$ are consistent for $b < 1$, but for the non–fully Bayes prior over models $\pi(M|\theta)$, $0 < \theta < 1$, the Bayesian procedure for $B_{j0}$ is inconsistent for $b = 1/2$ while the other two are start inconsistent. For $b = 1$ the Bayes factor $B_{j0}$ has to be ruled out of the inference and also the non–fully Bayes $\pi(M|\theta)$, $0 < \theta < 1$. A second conclusion is that as far as consistency is concerned, the hierarchal uniform prior outperforms the independent Bernoulli prior, and consequently the uniform prior on all models.

When sampling from an alternative model $M_T \neq M_0$ and $b = 1$, the Bayes factors are not consistent but there is a small region of alternative models around the null for which inconsistency holds. However, we have shown that the pairwise inconsistency region of the Bayes factor for the mixture of the $g-$priors strictly contains the inconsistency region of the Bayes factor for the intrinsic priors. In particular, we showed that when $\lim_{n\rightarrow\infty} n/k = 1$ the Bayes factor for the mixture of the $g-$priors is inconsistent under any alternative model.

The asymptotic analysis seems to support the conclusion that when the number of regressors grows with the sample size the Bayesian variable selection procedure based on intrinsic priors is preferred to those based on the mixture of $g-$priors.

Further, the bad asymptotic behavior of the non–fully Bayes Bayesian procedures we have considered has been corrected by considering a fully Bayes version of them. This conclusion is in agreement with that obtained by Scott and Berger (2010) who analyzed the empirical Bayes approach to estimating the hyperparameter $\theta$ of the independent Bernoulli prior for finite sample sizes.

**Acknowledgements** Authors partially supported by the grants: SEJ–02814 (Junta de Andalucía, Spain), ECO2009–14152 and MTM2011–28945 (Ministerio de Ciencia e Innovación, Spain).

## References

1. Casella, G. and Moreno, E.: Objective Bayesian variable selection. J. Amer. Statist. Assoc. **101**, 157–167 (2006)
2. Girón, F. J., Martínez, M. L., Moreno E. and Torres, F.: Objective Testing Procedures in Linear Models: Calibration of the $p-$values. Scandinavian Journal of Statistics **33**, 765–784 (2006)
3. Moreno, E., Girón, F. J. and Casella, G.: Consistency of objective Bayes factors as the model dimension grows. Ann. Statist. **38**, 1937–1952 (2010)
4. Scott, J. O. and Berger, J. O.: Bayes and empirical–Bayes multiplicity adjustment in the variable–selection problem. Ann. Statist. **38**, 2587–2619 (2010)
5. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. **6**, 461–464 (1978)