Comparison of Bayesian Objective Procedures for Variable Selection in Linear Regression

Elías Moreno and F. Javier Girón
Universidad de Granada and Universidad de Málaga

Abstract

In the objective Bayesian approach to variable selection in regression a crucial point is the encompassing of the underlying nonnested linear models. Once the models have been encompassed one can define objective priors for the multiple testing problem involved in the variable selection problem.

There are two natural ways of encompassing: one way is to encompass all models into the model containing all possible regressors, and the other one is to encompass the model containing the intercept only into any other.

In this paper we compare the variable selection procedures that result from each of the two mentioned ways of encompassing by analysing their theoretical properties and their behavior in simulated and real data.

Relations with frequentist criteria for model selection such as those based on the R^2_{adj}, and Mallows C_p are provided incidentally.

Keywords: *encompassing, intrinsic priors, linear regression, model selection, reference priors.*

1 Introduction

Suppose that Y represents an observable random variable and $X_1, X_2, \ldots, X_k$ a set of k potential explanatory covariates through the normal linear model

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + \varepsilon, \quad \varepsilon \backsim N(\cdot|0, \sigma^2). \tag{1}$$

An important problem, known as the variable selection problem, consists of reducing the complexity of model (1) by identifying a subset of the α_i coefficients that have a zero value based on an available dataset $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ is a vector of size n and $\mathbf{X}$ a $n \times k$ design matrix of full rank. This is essentially a model selection problem where we have to choose a model among the 2^k possible submodels

of the full model (1). It is common to set $X_1 = 1$ and $\alpha_1 \neq 0$ to include the intercept in any model. In this case the number of possible submodels is 2^{k-1}.

Subjective Bayesian variable selection has a long history, having been considered by Atkinson (1978), Smith and Spiegelhalter (1980), Pericchi (1984), Poirier (1985), Box and Meyer (1986), George and McCulloch (1993, 1995, 1997), Clyde, DeSimone and Parmigiani (1996), Geweke (1996), Smith and Kohn (1996), Kuo and Mallick (1998)

Typically, $0 - 1$ latent variables describing when each regression coefficient is zero are introduced. Then, the prior for the latent variables is a product of Bernoulli distributions. For linking the regression coefficients and the error variance with the latent variables standard prior distributions are considered, i.e., multivariate normal for the regression coefficients and inverse gamma for the error variance. The normal prior is centered at zero and the covariance matrix and the hyperparameters in the inverse gamma are often fixed with the help of some subjective/empirical criteria. Then, the posterior probabilities of the latent variables are computed and the covariates are selected as follows: a big enough posterior probability that a component of the latent variable is one implies to include the corresponding covariate as an explanatory variable.

This formulation involves a large number of hyperperparameters for which subjective inputs are required. Unfortunately, in the variable selection context subjective prior information on the regression coefficients and the variance errors of the models is typically not available.

Some attempts for solving the problem in a form "as objective as possible" are found in Mitchell and Beauchamp (1988) and Spiegelhalter and Smith (1982). In Mitchell and Beauchamp, the regression coefficients were assumed to be independent *a priori* and identically distributed with a prior distribution that concentrates some probability mass at zero with the rest of the prior mass spread out uniformly on a compact set. Conventional improper priors were used for the error variance. In their paper the variable selection problem is essentially an estimation problem that avoids the difficulties with improper priors; however, in order to specify the point masses some criteria are needed. Spiegelhalter and Smith used conventional improper priors for the regression coefficients and the variance error to compute Bayes factors and model posterior probabilities, but the arbitrary constants involved in the improper priors are determined by using subjective information on the marginal densities at some points of the sample.

A fully objective analysis for model comparison in linear regression was given in Berger and Pericchi (1996). They utilize an encompassing approach and an empirical measure, the intrinsic Bayes factor, which does not need subjective prior information. For large sample sizes this empirical measure closely approximates a Bayes factor, the Bayes factor for intrinsic priors. Variable selection procedures which use intrinsic priors have been considered by Casella and Moreno (2005a), and Girón et al. (2005a, 2005b).

Two Bayesian methods for variable selection that use intrinsic priors and different forms of encompassing linear models are compared in this paper. A first Bayesian procedure consists of considering the pairwise model comparison between the full model M_k and a generic submodel M_i having i ($< k$) nonzero regression coefficients. Since M_i is nested into the full M_k, this makes possible the derivation of intrinsic priors. Then, in the space of models $\{M_i, M_k\}$ the intrinsic posterior probability of M_i is computed and by doing so for all models M_i, $i = 1, \ldots, k$, an ordering in the set of all models $\mathcal{M}$ in accordance to their posterior probabilities $\{P(M_i|\mathbf{y}, \mathbf{X})$ is obtained. The interpretation of these probabilities is that the submodel having the highest posterior probability is the most plausible reduction in complexity from the full model, the second highest the second most plausible reduction and so on. This was the method followed by Casella and Moreno (2005a) and Girón et al. (2005a) and will be called in the sequel *variable selection from above*.

A second Bayesian procedure consists of considering the pairwise model comparison between a generic submodel M_i and the intercept only model

$$Y = \alpha_1 + \varepsilon, \quad \varepsilon \backsim N(\cdot|0, \sigma^2),$$

which is denoted as M_1. Notice that M_1 is nested in M_i, for any i, so that the corresponding intrinsic priors can be derived. In the space of models $\{M_1, M_i\}$ the intrinsic posterior probability $P^*(M_i|\mathbf{y}, \mathbf{X})$ is computed and it provides a new ordering of all the models in $\mathcal{M}$. This strategy was considered by Girón et al. (2005b) and will be called *variable selection from below*.

Both procedures may produce different ordering of the models and it is not clear which one of the two, if any, is preferred. The variable selection selection from above is based on multiple pairwise comparisons. This implies that for ranking the models intrinsic model posterior probabilities coming from different probability spaces, are being compared and one can argue that this might be not coherent. However, we note that all the posterior probabilities are computed with respect to the full model M_k and, consequently, each one indicates how plausible is a given reduction in complexity compared to the full model.

The variable selection selection from below is also based on multiple pairwise comparisons. However, we will see that it is equivalent to ordering the models according to intrinsic model posterior probabilities computed in the space of all models. Therefore, the possible lack of coherence of the variable selection from above approach is not shared by the variable selection from below procedure.

The paper is organized as follows. Section 2 gives the general form of the intrinsic prior distribution for any two nested normal linear models. Section 3 develops a general formula for computing the Bayes factor and the model posterior probabilities, and proves the consistency of the Bayesian procedure. The from above and from below Bayesian procedures are brevely reviewed in this section, and some of their known properties are revised and some new ones are

given. The relation among intrinsic model posterior probabilities, $p-values$, and the R^2 statistic is also considered. Section 4 discusses the role of encompassing in the Bayesian approaches and gives some clues about where the difficulties of the different criteria for variables selection arise. Section 5 is a short note on stochastic search. Section 6 contains comparisons of the numerical results obtained from the frequentist and Bayesian variable selection procedures for simulated and real datasets. Finally, section 7 gives some concluding remarks and recommendations.

2 Intrinsic priors for selecting between two nested linear models

Suppose we want to choose between the following two linear models

$$M_i : \mathbf{y} = \mathbf{X}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N_n(\mathbf{0}, \sigma_i^2\mathbf{I}_n),$$

and

$$M_j : \mathbf{y} = \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \quad \boldsymbol{\varepsilon}_j \sim N_n(\mathbf{0}, \sigma_i^2\mathbf{I}_n),$$

where M_i is nested in M_j. This implies that the $n \times i$ design matrix $\mathbf{X}_i$ is a submatrix of the $n \times j$ design matrix $\mathbf{X}_j$, so that $\mathbf{X}_j = (\mathbf{X}_i|\mathbf{Z}_{ij})$. Then, model M_j can be written as

$$M_j : \mathbf{y} = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{ij}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_j, \quad \boldsymbol{\varepsilon}_j \sim N_n(\mathbf{0}, \sigma_i^2\mathbf{I}_n).$$

Comparing these two models is equivalent to testing $H_0 : \boldsymbol{\beta}_0 = \mathbf{0}$ against $H_1 : \boldsymbol{\beta}_0 \neq \mathbf{0}$.

A Bayesian setting for this problem is that of choosing between the Bayesian models

$$M_i : N_n(\mathbf{y}|\mathbf{X}_i\boldsymbol{\alpha}_i, \sigma_i^2\mathbf{I}_n),\ \pi^N(\boldsymbol{\alpha}_i, \sigma_i) = \frac{c_i}{\sigma_i}, \tag{2}$$

and

$$M_j : N_n(\mathbf{y}|\mathbf{X}_j\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}_n),\ \pi^N(\boldsymbol{\beta}_j, \sigma_j) = \frac{c_j}{\sigma_j}, \tag{3}$$

where π^N denotes the improper reference prior (Berger and Bernardo, 1992).

The direct use of improper priors for computing model posterior probabilities is not possible; however, they can be converted into suitable intrinsic priors (Berger and Pericchi 1996, Moreno *et al.* 1998). Intrinsic priors for the parameters of the above nested linear models provide a Bayes factor (Moreno *et al.* 1998), and, more importantly, posterior probabilities for the models M_i and M_j, assuming that prior probabilities are assigned to them. An objective assessment of this prior is $P(M_i) = P(M_j) = 1/2$.

Application of the standard intrinsic prior methodology renders the intrinsic prior distribution for the parameters $(\boldsymbol{\beta}_j, \sigma_j)$ of model M_j, conditional on a fixed

parameter point $(\boldsymbol{\alpha}_i, \sigma_i)$ of the reduced model M_i,

$$\pi^I(\boldsymbol{\beta}_j, \sigma_j|\boldsymbol{\alpha}_i, \sigma_i) = \frac{2}{\pi\sigma_i(1+\sigma_j^2/\sigma_i^2)} N_j(\boldsymbol{\beta}_j|\tilde{\boldsymbol{\alpha}}_i, (\sigma_i^2+\sigma_j^2)\mathbf{W}_j^{-1}), \tag{4}$$

where $\tilde{\boldsymbol{\alpha}}_i^t = (\mathbf{0}^t, \boldsymbol{\alpha}_i^t)$ with $\mathbf{0}$ being the null vector of $j-i$ components. For assessing the matrix $\mathbf{W}_j^{-1}$ two close related forms have been proposed (Casella and Moreno 2005, Girón *et al.* 2005a). Although both essentially give a similar posterior answer, the computational simpler form is that given in Girón *et al.* (2005a) as

$$\mathbf{W}_j^{-1} = \frac{n}{j+1}(\mathbf{X}_j^t\mathbf{X}_j)^{-1},$$

which resembles the covariance matrix of Zellner's g-prior. Notice that the mean of the conditional intrinsic prior for $\boldsymbol{\beta}_j$ depends on the reduced model M_i but the matrix $\mathbf{W}_j^{-1}$ depends on the design matrix of the larger model $\mathbf{X}_j$. The reduced model M_i plays the role of the null hypothesis and the intrinsic prior for the parameters of the larger model M_j is "centered" at the null, a condition required in testing scenarios (Morris 1987; Casella and Moreno 2005a, 2005b). The marginal distribution of $\boldsymbol{\beta}_j|\boldsymbol{\alpha}_i$ is a proper prior that does not have moments, a condition required by Jeffreys when testing that the mean of a normal distribution is zero. The marginal distribution of the nuisance parameter $\sigma_j|\sigma_i$ is also a proper prior which does not have moments.

The unconditional intrinsic prior for $(\boldsymbol{\beta}_j, \sigma_j)$ is obtained from

$$\pi^I(\boldsymbol{\beta}_j, \sigma_j) = \int \pi^I(\boldsymbol{\beta}_j, \sigma_j|\boldsymbol{\alpha}_i, \sigma_i)\pi^N(\boldsymbol{\alpha}_i, \sigma_i)\, d\boldsymbol{\alpha}_i\, d\sigma_i. \tag{5}$$

Thus, the intrinsic priors for comparing models (2) and (3) are $\{\pi^N(\boldsymbol{\alpha}_i, \sigma_i), \pi^I(\boldsymbol{\beta}_j, \sigma_j)\}$.

3 Bayes factors and model posterior probabilities for nested models

Adapting the proof in Moreno et al. (2003), the Bayes factor to compare models M_i and M_j using the intrinsic priors $\{\pi^N(\boldsymbol{\alpha}_i, \sigma_i), \pi^I(\boldsymbol{\beta}_j, \sigma_j)\}$ turns out to be

$$B_{ij}(n, \mathcal{B}_{ij}) = \left(\frac{2(j+1)^{(j-i)/2}}{\pi} \int_0^{\pi/2} \frac{(\sin\varphi)^{j-i}(n+(j+1)\sin^2\varphi)^{(n-j)/2}}{(n\mathcal{B}_{ij}+(j+1)\sin^2\varphi)^{(n-i)/2}}\, d\varphi \right)^{-1}, \tag{6}$$

where the statistic $\mathcal{B}_{ij}$ is the ratio of two quadratic forms

$$\mathcal{B}_{ij} = \frac{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_j)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_i)\mathbf{y}} = \frac{SC_j}{SC_i}, \tag{7}$$

with $\mathbf{H}_i = \mathbf{X}_i(\mathbf{X}_i^t\mathbf{X}_i)^{-1}\mathbf{X}_i^t$, and $\mathbf{H}_j = \mathbf{X}_j(\mathbf{X}_j^t\mathbf{X}_j)^{-1}\mathbf{X}_j^t$ denoting the hat matrices of models M_i and M_j, and SC_i, and SC_j the residual sum of squares, respectively.

Thus, the intrinsic posterior probability of model M_i, when compared with M_j, is given by

$$P(M_i|n, \mathcal{B}_{ij}) = \frac{B_{ij}(n, \mathcal{B}_{ij})}{1 + B_{ij}(n, \mathcal{B}_{ij})}. \tag{8}$$

We note that formulae (7) and (8) make sense when $i < j$. However, we will adopt the convention that for $i = j$, $P(M_i|n, \mathcal{B}_{ii}) = 1/2$ since $\mathcal{B}_{ii} = 1$.

The asymptotic behavior of this posterior probability is as good as can be expected from a Bayesian procedure. Under mild conditions, the posterior probability $P(M_i|n, \mathcal{B}_{ij})$ tends, in probability, to 0 when sampling from model M_j, and to 1 when sampling from M_i. More precisely, adapting the proof of Moreno and Girn (2005c), we can prove the following theorem.

Theorem 1.
i) When sampling from model M_i we have

$$\lim_{n\to\infty} [P_i]\ P(M_i|n, \mathcal{B}_{ij}) = 1.$$

ii) If the limit

$$\mathbf{S}_{ij} = \lim_{n\to\infty} \frac{\mathbf{Z}_{ij}^t(\mathbf{I}_n - \mathbf{H}_i)\mathbf{Z}_{ij}}{n}$$

is a finite positive definite matrix, then when sampling from M_j

$$\lim_{n\to\infty} [P_j]\ P(M_i|n, \mathcal{B}_{ij}) = 0.$$

We remark that condition ii) to assure the consistency under the larger model is not a necessary but a sufficient condition. However, this is not a too demanding condition as the following example shows. This example has been taken from Berger and Pericchi (2004).

Example 1. Consider the case of testing whether the slope of a simple linear regression is zero. Thus, model M_1 is the model with only the intercept term α_1, and M_2 is the model with regression coefficients (α_1, α_2). Suppose there are $2n + 1$ observations and the design matrix is

$$\mathbf{X}^t = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 & 1 \\ 0 & \dots & 0 & \delta & \dots & \delta & 1 \end{pmatrix},$$

where δ is a real number. Easy calculations show that

$$\mathbf{S}_{12} = \frac{\delta^2}{4},$$

and hence, for $\delta \neq 0$ we have that $\mathbf{S}_{12}$ is a positive number.

This is a challenging example because, as n increases to infinity, we have an infinite number of samples coming from an alternative model which is as close to the null as we want by simply taking $|\delta|$ very small. However, the sufficient condition in ii) is satisfied; therefore, consistency holds.

3.1 Bayes factors and model posterior probabilities when M_k is the reference model

Under the first way of encompassing, which we can denote as *from above encompassing*, any generic model M_i with i regressors is nested to the full model M_k, so that the Bayes factor from formula (6) now becomes

$$B_{ik}(n, \mathcal{B}_{ik}) = \left(\frac{2(k+1)^{(k-i)/2}}{\pi} \int_0^{\pi/2} \frac{(\sin\varphi)^{k-i}(n+(k+1)\sin^2\varphi)^{(n-k)/2}}{(n\mathcal{B}_{ik}+(k+1)\sin^2\varphi)^{(n-i)/2}}\, d\varphi \right)^{-1},$$

where the statistics $\mathcal{B}_{ik}$ is the ratio of two quadratic forms

$$\mathcal{B}_{ik} = \frac{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H})\mathbf{y}}{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_i)\mathbf{y}} = \frac{SC}{SC_i},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ is the hat matrix of the full model M_k. This statistic is closely related to —is in fact a monotonic function of— the classical F statistic and the associated p-value for testing the null hypothesis M_i.

Thus, the intrinsic posterior probability of model M_i, when compared with M_k, is given by

$$P(M_i|n, \mathcal{B}_{ik}) = \frac{B_{ik}(n, \mathcal{B}_{ik})}{1 + B_{ik}(n, \mathcal{B}_{ik})}. \tag{9}$$

This pairwise probability computed for all models in $\mathcal{M}$ provide an ordering on the whole set of models.

This criterion has been extensively studied in Casella and Moreno (2005a) and Girón *et al.* (2005a). While the former paper focuses on variable selection, the latter contemplates its relation with the usual p-values.

3.2 Bayes factors and model posterior probabilities when M_1 is the reference model

Under the second way of encompassing, which we can denote as *from below encompassing*, the intercept only model M_1 is nested to any model M_i with i regressors, so that the Bayes factor from formula (6) becomes in this case

$$B^*_{1i}(n, \mathcal{B}^*_{1i}) = \left(\frac{2(i+1)^{(i-1)/2}}{\pi} \int_0^{\pi/2} \frac{(\sin\varphi)^{i-1}(n+(i+1)\sin^2\varphi)^{(n-i)/2}}{(n\mathcal{B}^*_{1i}+(i+1)\sin^2\varphi)^{(n-1)/2}}\, d\varphi \right)^{-1}, \tag{10}$$

where the statistics $\mathcal{B}^*_{1i}$ is given by

$$\mathcal{B}^*_{1i} = \frac{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_i)\mathbf{y}}{n\ s_y^2} = \frac{SC_i}{SC_1},$$

with $s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n$, and $\bar{y} = \sum_{i=1}^n y_i/n$.

Hence, the intrinsic posterior probability of M_i, when compared with M_1, is given by

$$P^*(M_i|n, \mathcal{B}^*_{1i}) = \frac{1}{1 + B^*_{1i}(n, \mathcal{B}^*_{1i})}. \tag{11}$$

The models M_i are now ordered according to their intrinsic posterior probabilities $\{P^*(M_i|n, \mathcal{B}^*_{1i})$ for all models in $\mathcal{M}$.

Note that the statistics $\mathcal{B}^*_{1i}$ can be expressed as

$$\mathcal{B}^*_{1i} = 1 - R^2_i, \tag{12}$$

where R^2_i represents the popular coefficient of determination, the ratio of the sum of squares due to regression of model M_i to the total sum of squares.

Observe that the intrinsic posterior probability $P^*(M_i|n, \mathcal{B}^*_{1i})$ given in (11) is a strictly increasing function of the Bayes factor $B^*_{i1}(n, \mathcal{B}^*_{1i})$ and, consequently, from (12) a strictly increasing function of R^2_i.

Remark 1. There is a close relation between the statistics involved in the computation of both Bayes factors. In fact, the product

$$\mathcal{B}_{ik} \cdot \mathcal{B}^*_{1i} = \frac{SC}{SC_1}$$

is a constant. However, there is no closed form relationship between the corresponding Bayes factors nor the pairwise posterior probabilities.

Unlike what happens in the from above encompassing procedure where the full model M_k needs to be fixed in advance, in the from below encompassing procedure the number of regressors to be included in the set of explanatory variables is not necessarily predetermined from the begining.

A nice property of the posterior probability given in (11) is that, if the number of explanatory variables i exceeds the sample size n, for instance when interactions are considered, then the posterior probability of *any model* with more regressors than the sample size is less that 1/2. This automatically prevents the consideration of explanatory models with too many regressors and, therefore, this procedure *always* penalizes complex models in accordance with Occam's razor principle. Theorem 2 proves this assertion, which is based on the following lemma.

Lemma 1. The upper bound of the Bayes factor (10), $\sup_{\mathcal{B}\in[0,1]} B^*_{1i}(n, \mathcal{B}) = B^*_{1i}(n, 1)$ is given by

$$\left(\frac{1}{\pi}\Big(\frac{n}{i+1}\Big)^{(n-i)/2} B\Big(\frac{1+i-n}{2}, \frac{1}{2}\Big) {}_2F_1\Big(\frac{i-n}{2}, \frac{1+i-n}{2}, \frac{2+i-n}{2}; -\frac{i+1}{n}\Big)\right)^{-1}. \tag{13}$$

Further, for all $i \geq n$, the sup is an increasing function of i.

Proof. As the Bayes factor is an increasing function of the statistic $\mathcal{B}$, the maximum is attained at $\mathcal{B} = 1$. For $\mathcal{B} = 1$, the Bayes factor (10) reduces to the expression

$$B^*_{1i}(n,1) = \left(\frac{2}{\pi}\int_0^{\pi/2}\left(1+\frac{n}{(i+1)\sin^2\varphi}\right)^{(n-i)/2}d\varphi\right)^{-1}.$$

This sup becomes 0 for $i < n$ as the integral diverges to infinity, and is equal to (13) for $i \geq n$. Note that for $i = n$, the sup equals 1. The proof that (*) is increasing in i, for $i \geq n$, is somewhat cumbersome as the proof uses some specific properties of the hypergeometric function and hence is is omitted.

Theorem 2. For any model M_i with i regressors such that $i \geq n$, and for all n, the posterior probability $P^*(M_i|n, \mathcal{B}^*_{1i})$ is a decreasing function of i that attains its maximum at $i = n$; the maximum of this probability being 1/2.

Proof. As $i \geq n$, the value of the R_i^2 statistic is equal to 0 and consequently $\mathcal{B}^*_{1i}$ is equal to 1; therefore, from lemma 1 and the fact that the posterior probability is a monotonic decreasing function of the Bayes factor, the theorem follows suit.

Very often, an interpretative difficulty of the probabilities given in (11) may arise when the number of covariates is large or even moderate. In fact, as we add new covariates the statistic R_i^2 increases rendering high values sometimes very close to one, so that the Bayes factor increases to infinity. In this case, and also for moderate or large sample sizes, many of the posterior probabilities $P^*(M_i|n, \mathcal{B}^*_{1i})$ tend to 1 making it difficult to interpret them as measures of discrimination for variable selection. Hald's data provide a nice example that illustrates this phenomenon.

These data are widely known and have been used as a benchmark for comparing variable selection criteria. They are interesting because the sample size $n = 13$ is small and the number of regression coefficients $k = 5$ is moderate, i.e., there are 4 covariates.

Table 1, which only includes the seven more probable models, shows the difficulties in interpreting the from below posterior probabilities as compared with model posterior probabilities exemplified, for example, in the two first rows of Table 1 where a difference of less than 10^{-6} in the from below posterior probabilities means an enormous difference in the model posterior probabilities computed in the space of all models that we now introduce as a way to remedy this difficulty.

Assuming the objective uniform model prior $P(M_i) = 1/2^{k-1}$, for $M_i \in \mathcal{M}$, the posterior probability of M_i in the space $\mathcal{M}$ is given by

$$\Pr(M_i|\mathbf{y},\mathbf{X})) = \frac{m_i(\mathbf{y},\mathbf{X})}{m_1(\mathbf{y},\mathbf{X}) + \sum_{j=2}^{2^{k-1}} m_j(\mathbf{y},\mathbf{X})}, \tag{13}$$

Models	From below Probs.	Model Probs.
$\{x_1, x_2\}$	0.999999	0.546571
$\{x_1, x_4\}$	0.999998	0.176554
$\{x_1, x_2, x_4\}$	0.999996	0.088912
$\{x_1, x_2, x_3\}$	0.999996	0.087916
$\{x_1, x_3, x_4\}$	0.999996	0.070828
$\{x_2, x_3, x_4\}$	0.999981	0.016538
$\{x_1, x_2, x_3, x_4\}$	0.999962	0.008199

Table 1: Comparison of pairwise and model posterior probabilities for Hald's data.

where $m_j(\mathbf{y}, \mathbf{X})$ represents the marginal of the data when using the intrinsic prior, i.e.,

$$m_j(\mathbf{y}, \mathbf{X}) = \int N_n(\mathbf{y}|\mathbf{X}_j\boldsymbol{\beta}_j, \sigma_j^2\mathbf{I}_n)\ \pi^I(\boldsymbol{\beta}_j, \sigma_j)\, d\boldsymbol{\beta}_j\, d\sigma_j,$$

and

$$m_1(\mathbf{y}, \mathbf{X}) = \int N_n(\mathbf{y}|\alpha_1\mathbf{1}_n, \sigma_1^2\mathbf{I}_n)\ \pi^N(\alpha_1, \sigma_1)\, d\alpha_1\, d\sigma_1.$$

Then, the posterior probability (13) can be written as

$$\Pr(M_i|\mathbf{y}, \mathbf{X}) = \frac{B^*_{1i}(n, \mathcal{B}^*_{1i})^{-1}}{1 + \sum_{j=2}^{2^{k-1}} B^*_{1j}(n, \mathcal{B}^*_{1j})^{-1}}. \tag{14}$$

We now observe that the posterior probability from below $P^*(M_i|n, \mathcal{B}^*_{1i})$ given in (11), obtained from the pairwise comparison, and that given in (14), which represents a probability in the space of all models, are both increasing functions of the Bayes factor $B^*_{i1}(n, \mathcal{B}^*_{1i})$. This implies that the ordering of the models given by the set $\{P^*(M_i|n, \mathcal{B}_{1i}),\ \ i \geq 1\}$ is exactly the same as that obtained from $\{\Pr(M_i|\mathbf{y}, \mathbf{X}),\, i \geq 1\}$. This last set of probabilities is a coherent set in the space of all models $\mathcal{M}$.

The use of the probabilities in the space of all models is useful *per se*, as they provide a coherent criterion for model selection and for model averaging, and allows for recognizing possibly masked differences in the paired-wise probabilities when these are too close to unity.

From now on, in the examples, we will only display model posterior probabilities instead of posterior probabilities from below.

Remark 2. Note that the ordering given by the posterior probabilities from below and the model probabilites is the same as far as we use objective uniform priors for both sets of models.

4 The role of encompassing

We have seen that the above two Bayesian procedures for variable selection essentially differ on the manner of encompassing the models. In the procedure from above the models are encompassed into the full, and in the procedure from below the intercept only model is encompassed into any possible model. It is interesting to analyze the role of encompassing to obtain some clues for the different orderings the procedures provide, and to relate them to frequentist criteria.

Let $\mathcal{M}_i$ denote the set of models of $\mathcal{M}$ that have exactly i regressors including the intercept. Thus, $\mathcal{M} = \bigcup_{i=1}^{k} \mathcal{M}_i$.

The frequentist *best subset regression* procedure chooses in each set $\mathcal{M}_i$ the model which minimizes the residual sum of squares

$$SC_i = \mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_i)\mathbf{y}.$$

In general, this produces a set of k *maximal* or *admissible* models each one with $1, \ldots, k$ covariates which always includes the intercept only model and the full model, as $\mathcal{M}_1 = \{M_1\}$ and $\mathcal{M}_k = \{M_k\}$. To choose among the maximal models from a frequentist viewpoint usually involves considering a trade-off between bias and variance; therefore, one has to use some additional criterion to order the set of maximal models.

In our Bayesian setting, from the expression of the statistic $\mathcal{B}_{ik}$ and formula (9) in section 3.1, it follows that under the from above procedure models in $\mathcal{M}_i$ are ordered in accordance with the values of the residual sum of squares SC_i obtained by considering all $\binom{k}{i-1}$ design matrices $\mathbf{X}_i$. The smaller the SC_i value the higher the Bayes factor. On the order hand, from the expression of the statistic $\mathcal{B}^*_{1i}$ in section 3.2 and formula (11), it follows that the ordering in $\mathcal{M}_i$ given by the from below procedure is also dictated by the values of the quadratic form SC_i.

Thus, within each $\mathcal{M}_i$ the ordering provided by the two Bayesian procedures is exactly the same. Therefore, both criteria always choose the maximal in each class $\mathcal{M}_i$ as does the *best subset regression* procedure. However, the ordering of the models on the whole class $\mathcal{M}$, when using either frequentist criteria or each of the two Bayesian procedures, does not necessarily coincide.

From the above discussion it is apparent that the differences resulting from the different criteria in the ordering of the models in the whole class $\mathcal{M}$ of models are due to the differences in dimension of the models.

From a frequentist viewpoint attempts to accomodate the different dimensions of the models have led to the adjusted version of the R^2 criterion defined as

$$R^2_{i,adj} = 1 - (1 - R^2_i)\Big(\frac{n-1}{n-i}\Big) = 1 - B^*_{1i}\Big(\frac{n-1}{n-i}\Big).$$

and to Mallows C_i

$$C_i = \frac{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H}_i)\mathbf{y}}{\mathbf{y}^t(\mathbf{I}_n - \mathbf{H})\mathbf{y}}(n-k) + (2i-n) = \frac{(n-k)}{\mathcal{B}_{ik}} + (2i-n), \quad 1 \leq i \leq k.$$

Bayesian procedures, on the other hand, automatically take into account the dimension of the model, when computing Bayes factors and model posterior probabilities.

Finally, we note that non exhaustive subset selection methods such as *forward stepwise selection*, *backward stepwise selection* and shrinkage methods as *ridge regression* and the *lasso* do not usually render maximal or admissible models.

5 Examples

The theoretical properties of the from above and from below Bayesian criteria do not provide conclusive evidence for deciding which of the two is best for model selection (see remark 1). They produce the same ordering when restricted to the classes $\mathcal{M}_i$ but they differ in the ordering of models with different number of covariates.

When model average is an important issue, as when prediction is involved, the from below procedure produces coherent posterior probabilities in the set of all contemplated models without need to specify a full model.

Thus, to test the performance of the from above and from below criteria, and to compare them with other frequentists criteria, we first consider how they behave against simulated data by measuring their respective performance in some specific way to be explained below. Later in this section, we make further comparisons based on two real data sets. Comparisons with the *lasso* procedure are only given for one of the well known real data sets described below as this procedure requieres splitting of the data and cross-validation.

5.1 Simulation study

The reported findings in this section are part of a larger simulation study carried out to compare the two Bayesian procedures. In general, both Bayesian procedures outperformed the adjusted R^2 and Mallows criteria for most simulations.

For the simulation study, we have considered small, medium and relatively large linear regression problems with sample sizes $n = 20$, $n = 40$, and $n = 100$, respectively, and with a small number of covariates; in particular set considered five covariates, i.e., $k = 6$.

The general linear model considered for simulation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is a vector ol length n, $\mathbf{X}$ is a $n \times 6$ matrix whose entries were obtained by simulation from a standard normal distribution $N(0,1)$, except the entries in

the first column which were set equal to 1 to include the intercept, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_6)^t)$ is a vector of length 6. The error terms coordinates of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^t$ are i.i.d. $\varepsilon_i \sim N(0, \sigma)$, where we set $\sigma = 1$ for all simulations.

After fixing $\mathbf{X}$, for the three sample sizes $n = 20, 40, 60$, 5000 samples of size n were simulated from the model for five different settings of the vector of regression coefficients $\boldsymbol{\alpha}(c)$ which was also varied according to a *tuning parameter* c, including 1, 2 3, 4 and 5 non zero coefficients. In particular, we set $c = 1, 2, 4, 8, 16$ and

$$\begin{aligned}
\boldsymbol{\alpha}_1(c) &= (2, -2, 0, 0, 0, 0)/c, \\
\boldsymbol{\alpha}_2(c) &= (2, -2, 2, 0, 0, 0)/c, \\
\boldsymbol{\alpha}_3(c) &= (2, -2, 2, -2, 0, 0)/c, \\
\boldsymbol{\alpha}_4(c) &= (2, -2, 2, -2, 2, 0)/c, \\
\boldsymbol{\alpha}_5(c) &= (2, -2, 2, -2, 2, -2)/c.
\end{aligned}$$

Note from our setting of the regression model that the simulation results depend not on the particular values of the regression coefficients α_i and the error standard deviation σ but on the value of α_i/σ and the sample size n.

The performance of all four criteria was measured by the proportion of times that the true model was selected in the first place in the 5000 simulations under the different values of n and the different settings for $\alpha_i(c)$.

For moderate and large samples, i.e. for $n = 40$ and $n = 100$, the Bayesian procedures dominate the frequentist criteria for most choices of the tuning parameter c, in the sense that they choose the true model more often than the frequentist procedures. For small values of c, i.e. when the covariates are more influential, the from below procedure dominates the from above one, whilst for large values of c, that is, when the regression coefficients are relatively less influential, the from above procedure seems slightly superior.

In general, both Bayesian procedures choose simpler models than the frequentist methods when the regression coefficients are relatively small, in accordance with Occam's razor principle. The from below procedure tends to choose even simpler models than the from above one when the tunig parameter c is large; thas means that the from below procedure applys Occam's principle in a more stringent way.

The from above procedure seems to be preferable for complex models, i.e. when the number of influential regressors is large, while the from *below* procedure seems to perform better for models where the number of influential variables is small or moderate.

For the small sample size simulations, $n = 20$, the conclusions are not so clear-cut, although the Bayesian criteria performed slightly better for small values of the tuning parameter c.

As expected, the adjusted R^2 criteria performed very poorly for almost all simulations —it was dominated by the other three criteria under all circunstances—, thus ruling it out as a model selection criterion. Mallows C_p, as it is based on the same statistic $\mathcal{B}_{ik}$ as the *from above* criterion, performed in parallel with the latter one but was otherwise dominated by both Bayesian procedures for most simulations except for those where the sample size was very small. In this case the behavior of Mallows criteria was similar to the Bayesian ones.

Similar conclusions were obtained for simulations carried out in models with a larger number of covariates.

5.2 Protein content

These data refer to a sample of size $n = 19$ of wheat yield and protein content measurements where the covariate is the wheat yield and the response is the protein content, and have been taken from Snedecor and Cochran (1982, p. 399). They fit a quadratic model to explain the protein content given the yield.

These data have also been analyzed by Guttman et al. (2004) assuming an unknown degree for the polynomial regression. They consider the model selection problem of choosing among polynomial models from the polynomial with degree zero up to the polynomial with degree d with d unknown. For estimating d, intrinsic Bayes factors, fractional Bayes factors are used and compared to more traditional model selection methods based on AIC, and BIC. The conclusion was that the posterior probability of the linear and quadratic polynomial support more than 0.90 of the posterior probability.

Here we also assume that the degree of the polynomial d is unknown. But instead of taking the set of d polynomial models as candidates for model selection, we consider the whole set of all 2^d possible polynomial models.

The two Bayesian procedures for $d \geq 3$ always select the same model with covariates x, x^3 followed closely for the quadratic polynomial including variables x, x^2 and a cubic model with no linear term. Interestingly, these three models provide nearly the same fitting for the data in the range of the x variable. The linear model, though simpler, always has smaller posterior probability than these three models whatever the value of $d \geq 2$. Table 2 displays the results of applying both Bayesian criteria to the whole class of eight polynomial models of degree up to $d = 3$.

The message implicit in our analysis is that either for the estimation of the order of the polynomial d or for the prediction of future values —from the viewpoint of model selection problems—, ought to include the set of *all* possible polynomials. Note that the full polynomial of degree 3, selected in the 5th place just behind the linear model, has a modest posterior probability of 0.072346. What it is important is the number of terms —not the degree— of the polynomial.

The conclusion is that a polynomial with a linear and cubic term provides a slihtly better model that the quadratic one and better than the linear model.

Polynomial Models	Model Probs.	Probs. From above
$\{1, x, x^3\}$	0.270729	0.768179
$\{1, x, x^2\}$	0.256300	0.758289
$\{1, x^2, x^3\}$	0.201814	0.712024
$\{1, x\}$	0.172328	0.572330
$\{1, x, x^2, x^3\}$	0.072346	0.500000
$\{1, x^2\}$	0.021229	0.174466
$\{1, x^3\}$	0.004827	0.057394
$\{1\}$	0.000428	0.005881

Table 2: Comparison of the two Bayesian criteria for the protein data

Consideration of polynomials of heigher order do not improve the model with two covariates.

5.3 Prostate cancer data

The data for this example come from a study by Stamey et al. (1989) and can be found in `http://www-stat.stanford.edu/ElemStatLearn`. These data are analysed, using different variable selection procedures, in Hastie et al. (2001). The response variable is the level of prostate-specific antigen (`lpsa`), and the eight covariates are the log cancer volume (`lcavol`), log prostate weight (`lweight`), `age`, log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores 4 or 5 (`pgg45`). The sample size is $n = 97$ and the number of regressors, including the intercept, is $k = 9$.

The set of admissible models turns out to be:

$$\{x_1\}$$
$$\{x_1, x_2\}$$
$$\{x_1, x_2, x_5\}$$
$$\{x_1, x_2, x_4, x_5\}$$
$$\{x_1, x_2, x_3, x_4, x_5\}$$
$$\{x_1, x_2, x_3, x_4, x_5, x_8\}$$
$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_8\}$$
$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$$

Using the whole data set, both the from below and from above procedures select the same model $\{x_1, x_2, x_5\}$, while the adjusted R^2 selects model $\{x_1, x_2, x_3, x_4, x_5, x_6, x_8\}$ and Mallows criterion selects model $\{x_1, x_2, x_4, x_5\}$.

We want to remark that our results are not wholy comparable to those of

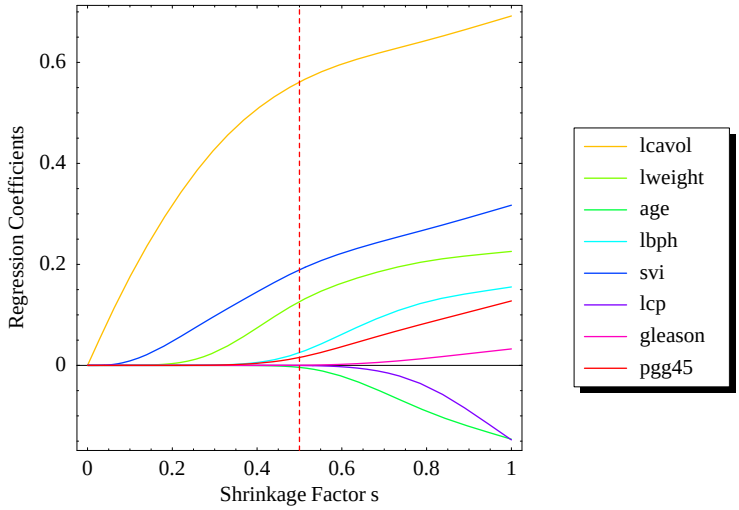

Figure 1: Profiles of the lasso coefficients for the prostate cancer data.

Hastie et al. (2001) as the latter analyse a training subset on size 67 leaving as a test set the remaining 30. Further, they apply 10-fold cross-validation to the training subset and find that the best subset criterion selects model $\{x_1, x_2\}$, while the lasso selects model $\{x_1, x_2, x_4, x_5, x_8\}$. Note that the lasso selects an inadmissible model of five variables if the shrinkage parameter is set at about $s \approx 0.5$. Our Figure 1, computed using all the data, resembles Figure 3.8 of Hastie et al. (2001), obtained using the training subset, but it is slightly different.

For these data all criteria choose models of different dimension. Except for the best subset criterion —that was chosen using a subset of size 67 — which selects the model with two covariates, namely `lacavol` and `lweight`, the two Bayesian criteria choose the most parsimonious model including the three covariates `lacavol`, `lweight` and `svi`.

6 Discussion

Objective Bayesian methodology seems to be particularly useful for the variable selection problem, a problem for which subjective prior information is typically not available and hence the use of objective priors is the natural alternative. To derive the objective priors for the multiple testing problem involved, an essential feature is that of encompassing the underlying models. Two different ways of ecompassing the models have been considered here: from *above* in which any model is nested into the full model, and from *below* in which the intercept only model is nested into any other. These two ways of encompassing generally produce different ordering of the models. On the other hand, under very mild conditions, both procedures are consistent (see, Moreno and Girón 2005c), so

that both select the true model as the sample size goes to infinity.

The from *below* and from *above* procedures depend on different but closesly related sufficient statistics as stated in remark 1, but the corresponding Bayes factors are only monotonously related for models with the same number of regressors. This implies that when ranking the models of the class $\mathcal{M}_i$, that have a fixed number of regressors i, the two Bayesian procedures, the adjusted R^2, and Mallows C_p render the same ranking as they too are also monote functions of the sufficent statistics. The ordering is dictated by minimazing the residual sum of squares SC_i: the smaller the SC_i the higher the posterior probability. However, when models with different number of regressors are considered the ordering differ. From the simulations, it follows that the Bayesian objective procedures turn out to be clearly superior to the frequentists adjusted R^2 and to Mallows C_p. This simply shows that to adjust for the dimensionality in variable selection is easy for any Bayesian procedure but a hard problem for the frequentists criteria.

We note that in the from *above* procedure the posterior probability of a model M_i is obtained in the space $\{M_i, M_k\}$, where k is the number of regressors deemed possible. On the other hand, the from *below* procedure provides model posterior probabilities in the space of all models $\mathcal{M}$, so that the resulting set of posterior probabilities are coherent. Further, we do not need to fix in advance the whole set of regressors. If fact, if we add a new regressor to the problem, the ordering of the previous models is not affected by the new regressor as long as uniform priors for models are used. A computational advantage of this fact is that we only need to compute the marginals of the new models arising by the inclusion of this new regressor and renormalize. Another important feature of the from below procedure is that models with a number of regressors greater that the sample size are automatically ruled out, as may happen when interactions among covariates are considered, thus acting as an automatic Occam's razor.

From the simulations performed it follows that both Bayesian procedures tend to choose simpler models than the frequentist methods, a sensible Occam's razor accomodation. The from *above* procedure seems to be preferable for complex models, i.e. when the number of influential regressors is large, while the from *below* procedure seems to perform better for models where the number of influential variables is small or moderate.

For the real examples considered both Bayesian procedures select essentially the same regressors, a selection that differ from those given either by the adjusted R^2, the Mallows C_p, or the *lasso* criteria.

We finally remark that along the paper we have avoided the important problem of computation by maintaining the number of regressors smaller than ten in all of our simulations. Certainly, the computation of either model posterior probabilities or even residual sum of squares are hard problems when the number of regressors is large or even moderate. In these cases some sort of stochastic search is absolutely necessary.

7 References

• Atkinson, A.C. (1978). Posterior probabilities for choosing a regression model. *Biometrika*, **65**, 39–48.

• Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds. Oxford University Press: Oxford, pp. 35–60.

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

• Berger, J.O. and Pericchi, L.R. (2004). Training samples in objective Bayesian model selection, *Annals of Statistics*, **32**, 841–869.

• Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorial. *Technometrics*, **28**, 11–18.

• Casella, G. and Moreno, E. (2005a). Objective Bayesian variable selection. *Journal of the American Statistical Association*, (to appear).

• Casella, G. and Moreno, E. (2005b). Intrinsic meta–analysis of contingency tables. *Statistics in Medicine*, **24**, 583–604.

• Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, **91**, 1197–1208.

• George, E. I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

• George, E. I. and McCulloch, R.E. (1995). Stochastic search variable selection. In *Practical Markov Chain Monte Carlo in Practice*, eds. W.R. Gilks et al. Chapman and Hall: London, 339–348.

• George, E. I., McCulloch, R.E. (1997). Approaches for variable selection. *Statistica Sinica*, **7**, 339–373.

• Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5*, eds. J. M. Bernardo et al., Oxford University Press: Oxford, 169–194.

• Giron, F.J., Martinez, M.L., Moreno, E. and Torres, F. (2005a). Objective testing procedures in linear models. Calibration of the p–values. Technical Report. University of Granada.

• Giron, F.J., Moreno, E. and Martinez, M.L. (2005b). An objective Bayesian procedure for variable selection in regression. Technical Report. University of Granada.

• Guttman, I., Peña, D. and Redondas, D. (2005). A Bayesain approach for predicting with polynomial regression of unknown degree. *Technometrics*, **47**, 23–33.

• Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical learning: Data Mining, Inference and Prediction*, Springer, New York.

• Kuo and Mallick (1998). Variable selection for regression models. *Sankhyā*, **60**, 65–81.

• Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, **83**, 1023–1036.

• Moreno, E., Bertolino, F. and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, **93**, 1451–1460.

• Moreno E., Girón, F.J. and Torres, F. (2003). Intrinsic priors for hypotesis testing in normal regression models. *Rev. R. Acad. Cien. Serie A, Mat., RACSAM*, **97** (1) 53–61.

• Moreno E. and Girón (2005c). Consistency of Bayes factors for intrinsic priors in normal linear models. *C. R. Acad. Sci. Paris, Ser. I*,

• Morris, C. M. (1987). Discussion of Casella and Berger. *Journal of the American Statistical Association*, **82**, 106–111.

• Pericchi, L. R. (1984). An alternative to the standard Bayesian procedure for discrimination between normal linear models. *Biometrika*, **71**, 575–586.

• Poirier, D. J. (1985).Bayesian hypothesis testing in linear models with continuously induced conjugate prior across hypotheses. In *Bayesian Statistics 2*, eds. J. M. Bernardo et al., Elsevier: New York, 711–722.

• Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–344.

• Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factor and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B*, **42**, 213–220.

• Snedecor, J. and Cochran, W. (1982). *Statistical Methods* (7th edition). Iowa State University Press.

• Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factor for linear and log–linear models with vague prior information. *Journal of the Royal Statistical Society, Series B*, **44**, 377–387.

• Stamey, T., Kabakin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II: radical prostatectomy treated patients. *Journal of Urology*, **16**, 1076–1083.