A Bayesian Approach for Predicting With Polynomial Regression of Unknown Degree

Irwin GUTTMAN

Department of Mathematics
State University of New York
Buffalo, NY 14260

Daniel PEÑA and Dolores REDONDAS

Department of Statistics
Universidad Carlos III de Madrid
Madrid 126, Getafe 2893, Spain
(*redondas@est-econ.es*)

This article compares three methods for computing the posterior probabilities of the possible orders in polynomial regression models. These posterior probabilities are used for forecasting using Bayesian model averaging. It is shown that Bayesian model averaging provides a closer relationship between the theoretical coverage of the high-density predictive interval (HDPI) and the observed coverage than those corresponding to selecting the best model. The performance of the different procedures is illustrated with simulations and some known engineering data.

KEY WORDS: Bayes information criterion; Bayesian model averaging; Fractional Bayes factor; Intrinsic Bayes factor.

1. INTRODUCTION

In many engineering situations where the response variable of interest is a polynomial function of an independent variable, an important problem is to determine the degree of the polynomial. From the frequentist standpoint, the most common approaches are (1) applying a variable selection method (e.g., forward or backward selection), which uses the t statistic for testing the coefficient of the highest-order polynomial, and (2) selecting the model by an order determination criterion, such as that of Akaike (1973) and others. From the Bayesian standpoint, two alternative options are available: (1) determining the order of the polynomial by means of the Bayes factors and (2) using an asymptotic approximation to the posterior model probabilities, such as the criteria of Schwarz (1978), Philips and Guttman (1998), and others.

Although these approaches are very useful for selecting the model that seems to have generated the data, they are less useful for forecasting purposes when there is a considerable uncertainty regarding the degree of the polynomial. In particular, the highest posterior prediction intervals, or the confidence intervals for the parameters, may be too short because the uncertainty about the degree of the polynomial involved is not completely taken into account. In this article we first compare different procedures for computing the posterior probabilities for different polynomial degrees, then take into account the model uncertainty for forecasting using Bayesian model averaging (BMA).

The main idea of BMA is as follows. Suppose that we have a set of possible models, $M_1, M_2, \ldots, M_K$, that can generate a given dataset $\mathbf{y}$. Suppose that we have prior probabilities, $P(M_i)$, and are able to compute the posterior probabilities of the models given the available data, $P(M_i|\mathbf{y})$. Then the predictive distribution of a new observation y_f can be obtained by weighting the predictive distributions of each model by their posterior probabilities, $P(M_i|\mathbf{y})$. Accordingly, BMA takes into account the uncertainty about the different models, as was pointed out in the seminal work of Leamer (1978). (See Draper and Guttman 1987; George 1999; Draper 1995; Chatfield 1995; Kass and Raftery 1995; Hoeting et al. 1999; Raftery et al. 1997; Fernandez et al. 2002 for different applications of this procedure.)

The probability $P(M_i|\mathbf{y})$ is proportional to $p(\mathbf{y}|M_i)P(M_i)$, and $p(\mathbf{y}|M_i)$ is obtained by averaging over the possible parameter values, which requires the posterior probabilities for the model parameters. If we do not have clear prior information about the parameters and want to use a reference or noninformative prior for them, then the probabilities $P(M_i|\mathbf{y})$ cannot be determined. To illustrate this problem, suppose that the model M_i depends on some parameter vector $\boldsymbol{\theta}_i$ and that the prior probabilities for these parameter vectors, $p(\boldsymbol{\theta}_i|M_i)$, are improper, that is, $p(\boldsymbol{\theta}_i|M_i) \propto g(\boldsymbol{\theta}_i)$, so that $p(\boldsymbol{\theta}_i|M_i) = c_i g(\boldsymbol{\theta}_i)$, which means that the integral of $g(\boldsymbol{\theta}_i)$ diverges. Then the marginal distribution of the data when M_i holds is given by

$$p(\mathbf{y}|M_i) = c_i \int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i) g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i,$$

and the posterior probability that model M_i holds is

$$p(M_i|\mathbf{y}) = c_i (m(\mathbf{y}))^{-1} \left\{ \int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i) g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i \right\} p(M_i), \quad (1)$$

where $m(\mathbf{y}) = \sum_{i=1}^{K} p(\mathbf{y}|M_i)p(M_i)$. Thus we see that this probability, which is needed for choosing among the models and for computing a forecast by BMA, depends on the unknown constant c_i. We note that, using (1) with the definition of $m(\mathbf{y})$ given below (1), $\sum_{i=1}^{K} p(M_i|\mathbf{y}) = 1$. The Bayes factor for comparing two models, M_i and M_j, is

$$B_{ij} = \frac{p(M_i|\mathbf{y})}{p(M_j|\mathbf{y})} = \frac{c_i}{c_j} \frac{p(\mathbf{y}|M_i)}{p(\mathbf{y}|M_j)} \frac{p(M_i)}{p(M_j)}, \quad (2)$$

and depends on the unknown and indeterminate ratio c_i/c_j.

Once this problem is solved, we can compute forecasts taking into account all sources of uncertainty as follows. For a given model M_i, the posterior predictive distribution, $p(y_f|\mathbf{y}, M_i)$

TECHNOMETRICS, FEBRUARY 2005, VOL. 47, NO. 1
DOI 10.1198/004017004000000581

when predicting a future observation, y_f, where we assume that y_f is independent of $\mathbf{y}$, is given by

$$p(y_f|\mathbf{y}, M_i) = \int p(y_f|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|\mathbf{y}, M_i)\, d\boldsymbol{\theta}_i, \tag{3}$$

where $p(\boldsymbol{\theta}_i|\mathbf{y}, M_i)$ is the posterior distribution for the parameters involved in model M_i. This predictive distribution takes into account the variability of the parameters, measured by $p(\boldsymbol{\theta}_i|\mathbf{y}, M_i)$. The unconditional predictive distribution is then found by

$$p(y_f|\mathbf{y}) = \sum_{k=1}^{K} p(y_f|\mathbf{y}, M_k)p(M_k|\mathbf{y}). \tag{4}$$

We use (4) in the sequel and refer to it as BMA, for indeed the predictive of y_f, given the data $\mathbf{y}$ stated in (4), is a weighting of predictives of y_f under models M_k, $k = 1, \ldots, K$, with the weights given by the posterior probabilities that model M_k holds.

This equation can also be written, inserting (3) in (4), as

$$p(y_f|\mathbf{y}) = \sum_{k=1}^{K} p(M_k|\mathbf{y}) \int p(y_f|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|\mathbf{y}, M_k)\, d\boldsymbol{\theta}_k,$$

which shows that by using BMA, we are taking into account both the parameter variability, as measured by the weighting over the possible parameter values made by the integral, and the model variability, as measured by the weighting over the possible models.

Here we focus on the general polynomial regression model, M_j,

$$y = \beta_0 + \beta_1 x + \cdots + \beta_j x^j + \epsilon,$$

where ϵ is $\mathrm{N}(0, \sigma^2)$ and the degree j is unknown but is assumed to be such that $0 \le j \le d$. To estimate j, a sample of values (x_i, y_i) are obtained for $i = 1, \ldots, n$. Thus for some j, the observations are generated by

$$\mathbf{y} = \mathbf{X}_j\beta_j + \epsilon, \tag{5}$$

where $\boldsymbol{\beta}_j = (\beta_0, \ldots, \beta_j)'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, and $\mathbf{X}_j = (\mathbf{1}, \mathbf{x}, \mathbf{x}^2, \ldots, \mathbf{x}^j)$, with the $n \times 1$ column vector $\mathbf{x}^k$ given by $\mathbf{x}^k = (x_1^k, \ldots, x_n^k)'$. Then, under model M_j,

$$E(\mathbf{y}|M_j) = \sum_{i=0}^{j} \beta_i \mathbf{x}^i, \qquad j = 0, 1, \ldots, d.$$

The rest of the article is organized as follows. Section 2 introduces three priors for the model space: one that is noninformative and two that favor the parsimony principle with respect to the degree of the polynomial. Section 3 presents three different approaches for computing the posterior probabilities of the models given the available data: the intrinsic Bayes factor (IBF) of Berger and Pericchi (1996b), the fractional Bayes factor (FBF) proposed by O'Hagan (1995), and an approximate method based on the Bayesian information criterion (BIC), proposed by Schwarz (1978). These methods are compared in a Monte Carlo study in Section 4 and using some real data examples in Section 5. Finally, Section 6 gives some concluding remarks.

2. THE PRIOR FOR THE MODELS

We consider three possible choices for the prior distribution $p(M_j)$. The first choice is the uniform distribution over the set of possible orders, $j = 0, 1, \ldots, d$, that is,

$$p(M_j) = (d + 1)^{-1}. \tag{6}$$

The second choice for $p(M_j)$ is a prior that penalizes the degree of the polynomial. We use a truncated geometric prior distribution over the degree of the polynomial,

$$p(M_j) = \frac{(1 - q)}{(1 - q^{d+1})} q^j, \qquad j = 0, 1, \ldots, d, \tag{7}$$

for $0 < q < 1$, where j is the degree of the model. We are interested in choosing a model, given the data, that is as parsimonious as possible, and with this aim, we have chosen the prior (7) that favors M_0, so that a priori $E(\mathbf{Y}) = \boldsymbol{\beta}_0$. Making a correspondence between J and M_j, this implies that we should choose the prior in such a way that $E(J) < .5$, that is, $E(J) = \frac{q}{(1-q)} \frac{1-(d+1)q^d+dq^{d+1}}{(1-q^{d+1})} < .5$, which, as may be verified, holds if we choose $q < 1/3$. The prior (7) decreases as j increases and has the advantage that the ratios $p(M_j)/p(M_{j+1})$ are constant for $j = 0, \ldots, d - 1$.

The third prior proposed is a truncated binomial prior distribution $B(n, p)$, which implies that

$$p(M_j) = cp^j(1 - p)^{n-j}, \qquad j = 0, 1, \ldots, d, \tag{8}$$

where c in chosen so that $\sum p(M_j) = 1$. We have chosen in the examples $n = 7$ and $p = 1/3$.

3. METHODS FOR DETERMINING POSTERIOR MODEL PROBABILITIES

Assuming the standard noninformative prior for $(\boldsymbol{\beta}_j, \sigma^2)$, to compute the posterior probabilities that model M_j holds (i.e., the degree is j) using (1), we would need the normalizing constant $(m(\mathbf{y}))^{-1}$. Using (1) with the definition of $m(\mathbf{y})$ that follows (1), we are involved with parameter vector $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i', \sigma^2)'$ of dimension $i + 2$ and also note that, using (2), it is straightforward to show that

$$p(M_j|D) = \frac{p(D|M_j)p(M_j)}{\sum_{i=0}^{d} p(D|M_i)p(M_i)} = \left[\sum_{i=0}^{d} B_{ij} \frac{p(M_i)}{p(M_j)}\right]^{-1}, \tag{9}$$

where $B_{ij} = p(D|M_i)/p(D|M_j)$ is the Bayes factor needed. It is important to note that when improper priors are used, the Bayes factors depend on the unknown indeterminate ratio c_i/c_j [see (2)]. Also, we use the notation D to denote the data $(\mathbf{X}_d, \mathbf{y})$ with the understanding that for model j, $j < d$, a subset of D is used, namely $(\mathbf{X}_j, \mathbf{y})$. We also remark that $\sum_{j=0}^{d} p(M_j|D) = 1$, which easily follows from (9).

3.1 Intrinsic Bayes Factors (IBF)

Berger and Pericchi (1996a, b) proposed solving the indetermination problem in the Bayes factor when using a noninforma-

tive prior for the parameters by selecting at random a training sample of minimum size and using this sample as data to compute a proper posterior distribution for the parameters. Then this posterior is used as prior for the analysis of the rest of the data. Of course, the result then may depend on the particular training sample used, and some kind of averaging is required to avoid this effect.

Let $\mathbf{X}_j$ be the $n \times (j+1)$ design matrix with the complete data and columns $(1, x, \ldots, x^j)$ used when fitting model M_j. Then a training sample of $m = j+2$ out of n observations is selected. We index the use of a particular training sample of size m by t, $t = 1, \ldots, T = \binom{n}{m}$, and we assume from now on that the first m observations of the $\mathbf{y}$ vector, say $\mathbf{y}_t$, and the first m rows of the matrix $\mathbf{X}_j$, say $\mathbf{X}_t(j)$, correspond to the training sample, $D_t = (\mathbf{X}_t(j), \mathbf{y}_t)$, and $D_{(-t)}$ refers to the rest of the data, say $(\mathbf{X}_{(-t)}(j), \mathbf{y}_{(-t)})$. Suppose that the standard noninformative prior for $(\boldsymbol{\beta}_j, \sigma^2)$ is used. Then the posterior distribution for the parameters given a training sample is

$$\begin{aligned} &p(\boldsymbol{\beta}_j, \sigma^2 | D_t) \\ &\quad = K_1 (\sigma^2)^{-(m/2+1)} \\ &\qquad \times \exp\left(-\frac{1}{2\sigma^2} \big(\mathbf{y}_t - \mathbf{X}_t(j)\boldsymbol{\beta}_j\big)' \big(\mathbf{y}_t - \mathbf{X}_t(j)\boldsymbol{\beta}_j\big) \right), \end{aligned} \quad (10)$$

where K_1 is a constant depending only on t. Now we use the posterior (10) as a prior for the remaining analysis. Note that

$$B_{ij} = \frac{p(D|M_i)}{p(D|M_j)} = \frac{p(D_{(-t)}|D_t, M_i) p(D_t|M_i)}{p(D_{(-t)}|D_t, M_j) p(D_t|M_j)} = B_{ij}(t) B_{ij}^t,$$

where $B_{ij}(t)$ is the conditional Bayes factor given the data in the training sample and B_{ij}^t is the Bayes factor using only the training sample. Thus we have that

$$B_{ij}(t) = B_{ij} B_{ji}^t.$$

Suppose that we use noninformative priors, so that B_{ij} and B_{ji}^t depend [as shown in (2)] on unknown constants. These constants will then be cancelled out when computing the conditional Bayes factor. Because the conditional Bayes factor depends on the training sample, Berger and Pericchi (1996a, b) proposed several types of averaging over all the possible training samples. One of their proposals is to use the arithmetic IBF, defined as the arithmetic mean of $B_{ij}(t)$ over the T possible training samples. This factor is very expensive to compute and is unstable for small sample sizes. A better solution is the expected IBF, say B_{ij}^E, which in nested models with $M_i \subset M_j$ can be computed by

$$\begin{aligned} B_{ij}^E = C_{ij}^* \frac{|\mathbf{X}_j'\mathbf{X}_j|^{1/2} |\mathbf{X}_t'(i)\mathbf{X}_t(i)|^{1/2}}{|\mathbf{X}_i'\mathbf{X}_i|^{1/2} |\mathbf{X}_t'(j)\mathbf{X}_t(j)|^{1/2}} \left(\frac{R_i}{R_j}\right)^{-(n-i-1)/2} \\ \times [\exp(-\lambda_{ij}(t)/2)] \mathbf{M}\left(\frac{1}{2}, \frac{j-i+1}{2}, \frac{\lambda_{ij}(t)}{2}\right), \end{aligned}$$

where $R_j = \mathbf{y}'(\mathbf{I}_n - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j')\mathbf{y}$ is the sum of squares of the residuals for $\mathbf{X}_j$,

$$C_{ij}^* = \frac{\Gamma(\frac{n-i-1}{2})\Gamma(\frac{i-j+1}{2})}{\Gamma(\frac{n-j-1}{2})\Gamma(\frac{1}{2})} \left(\Gamma\left(\frac{1}{2}\right)^{-1}\right) \left(\frac{n-i-1}{2}\right)^{(j-i)/2}, \quad (11)$$

$$\begin{aligned} \lambda_{ij}(t) = \frac{R_i}{n-j-1} \boldsymbol{\beta}_i' \mathbf{X}_t'(i) \\ \times \big[\mathbf{I} - \mathbf{X}_t(j)\big(\mathbf{X}_t'(j)\mathbf{X}_t(j)\big)^{-1}\mathbf{X}_t'(j)\big]^{-1} \mathbf{X}_t(i)\boldsymbol{\beta}_i, \end{aligned} \quad (12)$$

and $\mathbf{M}(a, b, c)$ is Kummer's function (see Abramowitz and Stegun 1970, chap. 13). Then Abramawitz and Stegun (1970) define $B_{ji}^E = 1/B_{ij}^E$. We use the expected IBF for the comparison of the posterior probabilities. The posterior probabilities can be obtained using (9) by

$$p_I(M_j|D) = \left(\sum_{i=0}^{d} B_{ij}^E \frac{p(M_i)}{p(M_j)} \right)^{-1},$$

where the ratio $p(M_i)/p(M_j)$ depends on the particular prior for the models used.

3.2 Fractional Bayes Factor (FBF)

O'Hagan (1995) proposed avoiding the problem of indetermination with noninformative priors by using a modified Bayes factor, the FBF. For a dataset $D = (\mathbf{X}, \mathbf{y})$, this is defined as

$$B_{ij}^b(D) = \frac{q_i(b, D)}{q_j(b, D)}, \quad (13)$$

where $b = m/n$ and m is the size of the minimal training sample, with

$$q_i(b, D) = \frac{\int g(\boldsymbol{\theta}_i) p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, d\boldsymbol{\theta}_i}{\int g(\boldsymbol{\theta}_i) [p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)]^b\, d\boldsymbol{\theta}_i}, \quad (14)$$

where $g(\boldsymbol{\theta}_i)$ is the prior distribution for the parameters and $p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)$ is the full likelihood under the model M_i. Note that if $b = 0$, then no training sample is involved. Then (13) is just the standard Bayes factor, $B_{ij}(D) = B_{ij}^0(D)$, for comparing models M_i and M_j. The posterior probability for a model can now be written as

$$p(M_j|D) = \left[\sum_{i=0}^{d} B_{ij}^b \frac{p(M_i)}{p(M_j)} \right]^{-1}. \quad (15)$$

Now $q_i(b, D)$ may be computed for the polynomial model (5) using a noninformative prior for the parameters $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma^2)$, given by $g(\boldsymbol{\theta}_i) = p(\boldsymbol{\beta}_i, \sigma^2) \propto \sigma^{-2}$. Then the denominator of $q_i(b, D)$ in (14) is

$$\begin{aligned} &\int g(\boldsymbol{\theta}_i) [p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)]^b\, d\boldsymbol{\theta}_i \\ &\quad = (2\pi)^{-nb/2} \int \sigma^{-(nb/2+1)} \\ &\qquad \times \exp\left\{ -\frac{b}{2\sigma^2} \big(R_i + (\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)' \mathbf{X}_i'\mathbf{X}_i (\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i)\big) \right\} d\boldsymbol{\beta}_i\, d\sigma^2, \end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}$ and

$$R_i = \mathbf{y}'\big(\mathbf{I}_n - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\big)\mathbf{y}. \quad (16)$$

Integrating with respect to σ^2 and $\boldsymbol{\beta}_i$, we have that the denominator of $q_i(b, D)$ is

$$\begin{aligned} &\int g(\boldsymbol{\theta}_i) [p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)]^b\, d\boldsymbol{\theta}_i \\ &\qquad = \frac{1}{2} (\pi R_i)^{-w/2} b^{-nb/2} \Gamma\left(\frac{w}{2}\right) |\mathbf{X}_i'\mathbf{X}_i|^{-1/2}, \end{aligned}$$

where $w = nb - d - 1$ are the degrees of freedom. Because the numerator of $q_i(b, D)$ is equal to the expression for the denominator with $b = 1$, we then have that

$$q_i(b, \mathbf{y}) = \frac{\int g(\boldsymbol{\theta}_i) p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, d\boldsymbol{\theta}_i}{\int g(\boldsymbol{\theta}_i)[p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)]^b\, d\boldsymbol{\theta}_i} = \frac{\Gamma(\frac{v}{2})}{\Gamma(\frac{w}{2})} b^{+nb/2} (\pi R_i)^{-n(1-b)/2}.$$

For our polynomial problem, the minimum sample size that makes the prior proper for the parameters is $m = d + 2$, and then $b = m/n = (d + 2)/n$. To compute the posterior probability for the models, we use

$$B_{ki}^b(D) = \frac{\Gamma(\frac{n-k-1}{2})\Gamma(\frac{nb-i-1}{2})}{\Gamma(\frac{nb-k-1}{2})\Gamma(\frac{n-i-1}{2})} \left(\frac{R_k}{R_i}\right)^{-n(1-b)/2}, \qquad (17)$$

and from (15), the model posterior probabilities are

$$p(M_j|D) = K_{FBF} \frac{\Gamma(\frac{n-j-1}{2})}{\Gamma(\frac{d+1-j}{2})} (R_j)^{-(n-d-2)/2},$$

where

$$K_{FBF} = \left[\sum_{i=0}^{d} \frac{\Gamma(\frac{n-i-1}{2})}{\Gamma(\frac{d+1-i}{2})} (R_i)^{-(n-d-2)/2}\right]^{-1}.$$

3.3 The Bayes Information Criterion (BIC) Approximation

An alternative approach is to compute the posterior probabilities $p(M_j|D)$ using the BIC approximation. The Schwarz criterion for M_i is defined as

$$S(M_i) = \log p_i(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i) - \frac{1}{2} d_i \log n,$$

where $\widehat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimator (MLE) of the parameter vector $(\boldsymbol{\beta}_i, \sigma)$ under model M_i and d_i is the dimension of the vector $\boldsymbol{\beta}_i$. The BIC of a model M_i is

$$BIC(M_i) = -2S(M_i),$$

and, as Kass and Raftery (1995) pointed out, $\exp(S(M_i) - S(M_j))$ approximates the Bayes factor B_{ij} with a relative error $O(1)$. Then we can approximate the Bayes factors by

$$B_{ij}^{BIC} = \exp(S(M_i) - S(M_j)) = \frac{\exp(-.5BIC(M_i))}{\exp(-.5BIC(M_j))}$$

and obtain the posterior probability for a model by

$$p(M_j|D) \propto p(M_j) \exp\left(\log p_j(\mathbf{y}|\widehat{\boldsymbol{\theta}}_j) - \frac{1}{2} d_j \log n\right).$$

The likelihood for a normal linear model evaluated at the MLE $\widehat{\boldsymbol{\theta}}_j$ of $(\boldsymbol{\beta}_j, \sigma)$ is easily seen to be

$$p_j(\mathbf{y}|\widehat{\boldsymbol{\theta}}_j) = (2\pi)^{-n/2} \left(\frac{R_j}{n}\right)^{-n/2} e^{-n/2},$$

and the posterior probability of M_j, may be approximated, after absorbing common constants, by

$$p(M_j|D) = K_{BIC} p(M_j) R_j^{-n/2} n^{-(j+1)/2}, \qquad (18)$$

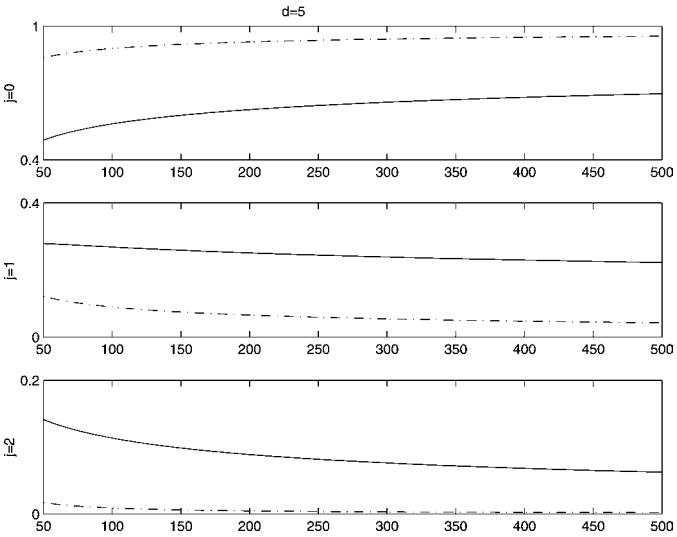

Figure 1. Standardized Penalty of the BIC Method (-···) and FBF Method (—). The maximum degree d is 5, for the models, constant, j = 0; linear, j = 1; and quadratic j = 2.

where

$$K_{BIC} = \left[\sum_{j=0}^{d} R_j^{-n/2} n^{-(j+2)/2} p(M_j)\right]^{-1}.$$

It is interesting to compare (18) with the probabilities obtained by the previous methods. It can be shown (see the App.) that assuming a uniform prior, the posterior probabilities computed by the FBF and the BIC approximation can be written as a function of the residual sum of squares and a penalty function that depends on the order of the polynomial. Figure 1 shows these standardized penalty functions as functions of the sample size, n, for $j = 0, 1, 2$ with maximum degree $d = 5$, where n is allowed to vary between 50 and 500. Similar results were obtained for other d. This figure shows that BIC penalizes more than FBF; BIC gives more weight than FBF to the model with lowest degree, and thus gives less weight than FBF to polynomials of higher degree.

4. SIMULATIONS

In this section we compare, using a Monte Carlo study, the procedures presented in the previous section. We also study the effect of the prior distribution on these procedures using the three priors defined in Section 2. We envision the following scenario: We generate observations using a model M_j of Table 1 (see Fig. 2 for a sample generated from each one of

Table 1. Model Used in the Monte Carlo Study

Model	$y =$	σ^2
M_1	$2 + x + \varepsilon$	1
M_2	$3 - x + \varepsilon$	1
M_3	$10 - 2x^2 + \varepsilon$	5
M_4	$-10 - 3x + x^2 + \varepsilon$	5
M_5	$3 + 10x - 2x^3 + \varepsilon$	10
M_6	$-4 + x - 3x^2 + x^3 + \varepsilon$	10

NOTE: In all the six cases, the distribution of the error term is N(0, σ^2).

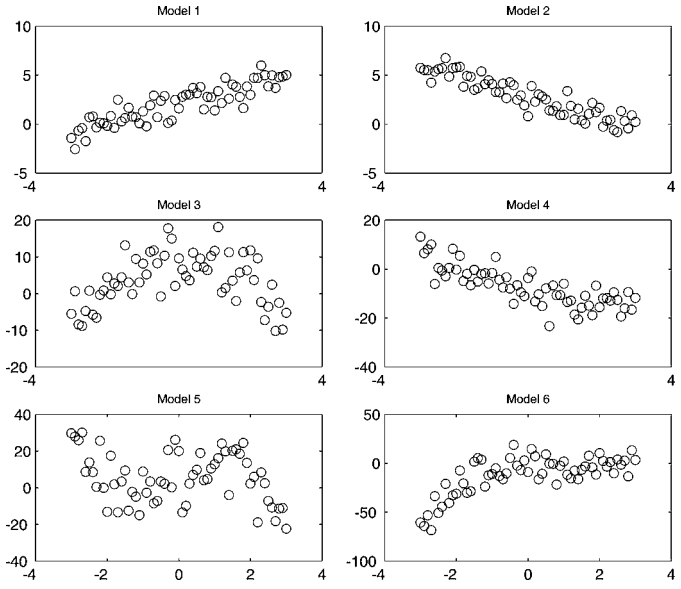

Figure 2. One Replication for the Six Models Under Study.

these models), where the x values are spaced equally in the interval $[-3, 3]$, so that the sample size in each case is $n = 61$. We generate $N = 100$ replications from each model. For each replication, we fit polynomial models of order $0, 1, \ldots, d = p + h$, where p is the *correct degree* of each model and $h = 0, 1, 2, 3, 4$, and compute the posterior probabilities for each possible order using the three methods described in Section 3 and the three priors described in Section 2.

4.1 Posterior Probabilities of the Models

Table 2 gives the posterior probabilities that the model M_i holds obtained using the nine methods. We emphasize (highlighting in bold type) the maximal posterior probability for each model and for each value of h.

We note that the posterior probability that M_i holds, given the data D, represents an updating of information. We start with prior information that M_i holds, namely $p(M_i)$, and then take into account the relevant data D and proceed as indicated in Sections 1–3 to update the prior, to the posterior that M_i holds, given D, using Bayes's theorem, which produces $p(M_i|D)$. As seen generally in Section 1, and in particular in (9), we have that $\sum_{i=0}^{d} p(M_i|D) = 1$.

An experiment confronted with the set of these posterior probabilities $p(M_i|D)$, $i = 0, \ldots, d$, might well look for the maximum, say

$$p(M_k|D) = \max_i \{p(M_i|D)\},$$

and use this result to fit model M_k. We advocate taking this approach in this situation because, based on prior and sample

Table 2. Posterior Probability That the Model M_i Holds, Using the Nine Methods

	IBFG	IBFB	IBFU	BICG	BICB	BICU	FBFG	FBFB	FBFU
M_1									
$h=0$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$h=1$	**.9913**	.9374	.9284	.9726	.8286	.8054	.9788	.8553	.8348
$h=2$	**.9911**	.9255	.9109	.9692	.7973	.7528	.9663	.7739	.7283
$h=3$	**.9937**	.9376	.9221	.9754	.7963	.7377	.9653	.7236	.6514
$h=4$	**.9923**	.9308	.9159	.9720	.7985	.7479	.9565	.6922	.6122
M_2									
$h=0$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$h=1$	**.9936**	.9442	.9351	.9783	.8324	.8083	.9832	.8601	.8389
$h=2$	**.9921**	.9242	.9089	.9737	.7963	.7537	.9700	.7713	.7274
$h=3$	**.9938**	.9412	.9286	.9786	.8291	.7850	.9696	.7599	.7005
$h=4$	**.9904**	.9291	.9142	.9733	.8079	.7545	.9588	.7015	.6176
M_3									
$h=0$	.9983	1.0000	1.0000	.9998	1.0000	1.0000	.9996	1.0000	1.0000
$h=1$	**.9849**	.9300	.8877	.9807	.8906	.8094	**.9849**	.9112	.8412
$h=2$	**.9861**	.9292	.8696	.9760	.8716	.7584	.9727	.8539	.7328
$h=3$	**.9895**	.9369	.8747	.9782	.8775	.7496	.9688	.8285	.6653
$h=4$	**.9854**	.9281	.8601	.9738	.8635	.7439	.9587	.7902	.6066
M_4									
$h=0$	.7343	.9120	.9209	.8364	.9565	**.9618**	.7941	.9413	.9480
$h=1$	.7460	.8663	.8411	.8366	.8578	.7888	.8283	**.8753**	.8160
$h=2$	.7202	**.8456**	.8152	.7893	.8141	.7156	.7870	.8037	.6966
$h=3$	.7332	**.8637**	.8151	.8179	.8416	.7294	.8158	.8004	.6536
$h=4$	.7508	**.8590**	.8057	.8169	.8274	.7143	.8121	.7641	.5892
M_5									
$h=0$	.7058	.9787	.9861	.8424	.9979	**.9988**	.8073	.9965	.9981
$h=1$	.7164	.8769	.8067	.8249	.9104	.7913	.8306	**.9278**	.8240
$h=2$	.7806	**.9248**	.8330	.8282	.9199	.7703	.8398	.9110	.7463
$h=3$	.7879	**.9194**	.8101	.8233	.9091	.7328	.8371	.8818	.6525
$h=4$	.7739	**.9056**	.7770	.7858	.8994	.7276	.8039	.8539	.5950
M_6									
$h=0$	.4966	.6637	.7166	.6675	.8367	**.8829**	.6014	.7980	.8536
$h=1$	.3951	.5775	.6118	.5610	.7336	.6941	.5409	**.7397**	.7198
$h=2$	.4673	.6535	.6570	.6538	**.7764**	.6707	.6454	.7761	.6580
$h=3$	.4634	.6603	.6873	.6138	**.7576**	.6806	.6144	.7540	.6241
$h=4$	.4527	.6450	.6642	.6054	**.7355**	.6378	.6083	.7226	.5406

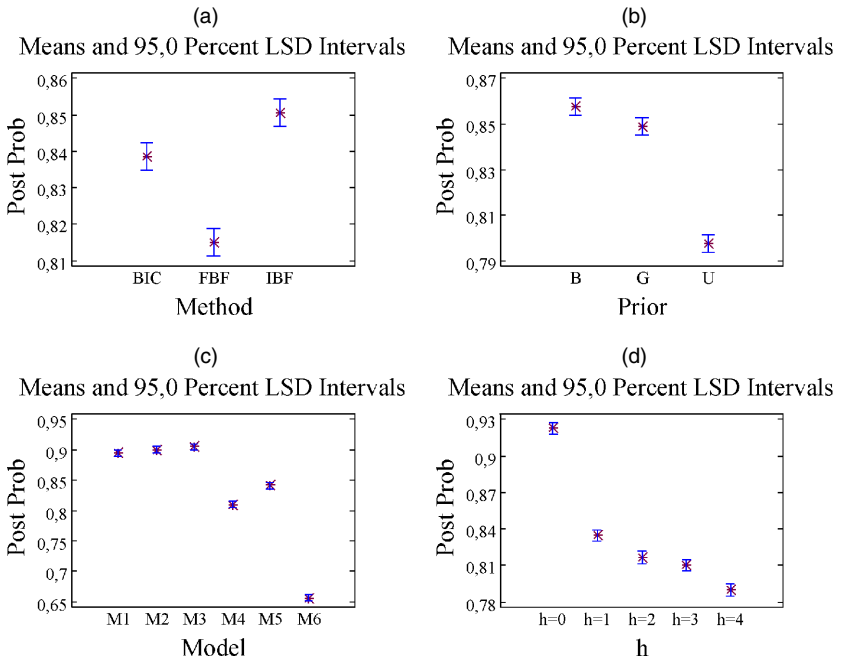

Figure 3. Main Effects and 95% Confidence Intervals for the Posterior Probabilities of a Correct Model.

information, the posterior probability that M_k holds is maximum, so that the experimenter has the most confidence in fitting model M_k and acting as if M_k is the model that best explains the data.

The results of Table 2 can be analyzed as an experiment with four factors: method, prior, model, and h. Figure 3 shows the estimation of the main effects for the four factors and the 95% confidence intervals for these estimations. We conclude that the IBF has the best performance, followed by BIC and FBF. Regarding the priors, the binomial is the best, followed closely by the geometric. The three first models have small uncertainty about the polynomial degree, and for them the probabilities of the correct degree are very high with all the methods, whereas the last three models, especially model M_6, show more uncertainty. The effect of increasing h is to decrease the posterior probabilities of a correct model, as expected.

Some interactions between priors and models are found. For the first three models (M_1–M_3), the probabilities are higher with the geometric prior, whereas the binomial is better for higher-order models (M_4–M_6). There are also interactions between models and methods. For the three first models, IBF is the best, whereas BIC is the best for models M_4–M_6. Regarding the interactions between priors and methods, IBF works better with the binomial, whereas BIC and FBF work better with the geometric. From the standpoint of robustness to the value of h, the more robust methods are BIC and IBF, and this important property, as well as their good overall performance, lead us to recommend them.

4.2 Predictive Distributions

To compare the prediction ability of the three methods, we generate observations at 10 equally spaced points in the interval $x_h = [-3, 3]$ and compute the response at these points from the six models described in Table 1. We also introduce in the comparison three standard methods for selecting the best model in linear regression applied to finding the order of the polynomial degree: all-subsets regression with the Akaike information criterion (AIC) (Akaike 1973), the *forward selection* (FS) method and the *backward elimination* (BE) method (see, e.g., Thompson 1978). The process is repeated 100 times, and the frequency with which the true values are included in the 85%, 90%, 95%, 97.5%, and 99% highest predictive density interval (HPDI) obtained by the 12 methods considered is recorded. The HPDI have been computed for the Bayesian methods by (a) BMA and (b) selecting the best model (SBM). For the three classical procedures, the prediction confidence intervals (PCIs) are those provided by the best model selected.

Let $\mathbf{f}(\boldsymbol{\alpha}, i)$, with $\boldsymbol{\alpha} = (.85, .90, .95, .975, .99)$, be the relative frequency with which the true value is included in the $HPDI(\boldsymbol{\alpha}, i)$ interval, i =IBFG, IBFB, IBFU, BICG, BICB, BICU, FBFG, FBFB, FBFU, or in the PCI$(\boldsymbol{\alpha}, i)$, with $i =$ AIC, FS, BE. Let

$$\mathbf{d}(\boldsymbol{\alpha}, i) = (\mathbf{f}(\boldsymbol{\alpha}, i) - \boldsymbol{\alpha})100$$

be the percentage deviation between the observed interval coverage and the theoretical interval coverage. Table 3 presents the values of $\mathbf{d}(\boldsymbol{\alpha}, i)$. The best result for each model is highlighted in bold type.

We can observe that all the values are negative, which indicates that all of the methods underestimate the length of the true predictive interval; that is, they underestimate the uncertainty involved in forecasting. For the three standard methods, FS works better in the first two methods, whereas AIC shows better behavior in the last three methods. Further examination of the results shows that BE does not perform very well, but that for M_5, the methods with a geometric prior demonstrate worse results than BE, and again for M_5, FS is worse than BE. We also note that prediction intervals generated by BMA almost always have better coverage than those generated by the best selected

Table 3. Results for the Mean of the Difference Between the Number of Points Contents in the α HDI and the Nominal Value α Multiplied by 100

	IBFG	IBFB	IBFU	BICG	BICB	BICU	FBFG	FBFB	FBFU	AIC	BE	FS
M_1												
BMA	−.0002	−0	.0003	−.0001	−.0003	−.0002	−.0002	−.0002	−.0002			
SBM	−.0002	−.0007	−.0008	−.0003	−.0011	−.0014	−.0003	−.0013	−.0015	−.0014	−.0039	−.0009
M_2												
BMA	−.0032	−.0036	−.0029	−.0035	−.0033	−.0030	−.0034	−.0030	**−.0028**			
SBM	−.0032	−.0036	−.0036	−.0036	−.0042	−.0043	−.0036	−.0040	−.0038	−.0044	−.0057	−.0035
M_3												
BMA	−.0036	−.0035	**−.0027**	−.0037	−.0035	−.0043	−.0037	−.0038	−.0045			
SBM	−.0039	−.0039	−.0042	−.0039	−.0048	−.0054	−.0039	−.0050	−.0058	−.0057	−.0147	−.0054
M_4												
BMA	−.0104	−.0103	**−.0094**	−.0110	−.0107	−.0116	−.0102	−.0109	−.0116			
SBM	−.0156	−.0127	−.0123	−.0140	−.0117	−.0126	−.0145	−.0118	−.0129	−.0126	−.0153	−.0144
M_5												
BMA	−.0170	**−.0153**	−.0154	−.0174	−.0160	−.0166	−.0171	−.0156	−.0165			
SBM	−.0303	−.0160	−.0171	−.0263	−.0165	−.0182	−.0238	−.0165	−.0183	−.0185	−.0209	−.0240
M_6												
BMA	−.0119	−.0110	**−.0088**	−.0130	−.0129	−.0158	−.0112	−.0127	−.0154			
SBM	−.0205	−.0195	−.0204	−.0200	−.0192	−.0200	−.0195	−.0192	−.0205	−.0201	−.0224	−.0206

model. Note that intervals by BMA are larger than those by SMB, but keep in mind that this property does not imply that they have better coverage.

The results of Table 3 can also be analyzed as an experimental design with four factors: method, prior, procedure for prediction (BMA or SBM), and model. To include the procedure as a factor, we eliminate the three non-Bayesian methods (AIC, FS, and BE) in the experimental design. Figure 4 shows the estimated main effects and their 95% confidence intervals. The best performance is again obtained by IBF, but now the differences among the three methods are not significant at the .05 level, corresponding to a 95% confidence interval. With regard to the priors, the binomial works better, but there is no significant difference from the uniform, although both have a significantly better performance than the geometric prior. There is a very pronounced effect of predicting using BMA with respect to SBM, with predictive coverage improved by approximately 30%. Finally, the uncertainty in the model is in agreement with previous results, although whereas the last model, M_6, was the one with the lowest posterior probability for the correct model, it is model M_5 that has the worst predictive capability. The conclusions with respect to the interactions are similar to those of the previous case.

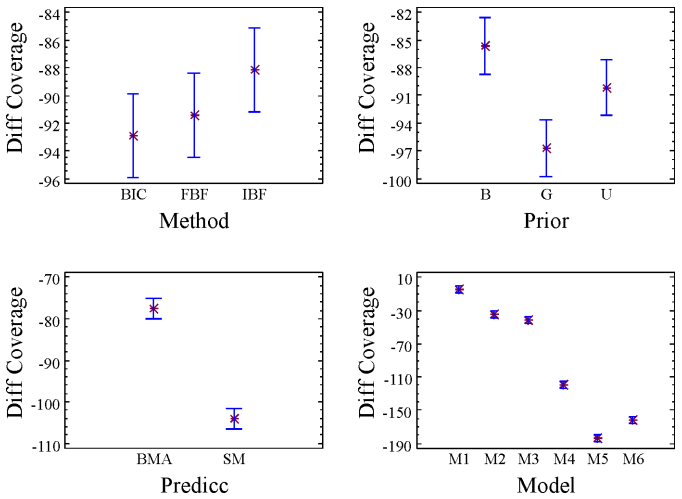

Figure 4. Main Effects and 95% Confidence Intervals for the Difference Between Coverage of the Interval and the Nominal Value. The scale is multiplied by 10^{-3}.

5. EXAMPLES

5.1 Protein Content

The data on wheat yield and protein content were reported by Snedecor and Cochran (1989, p. 399). This dataset has $n = 19$ and is presented graphically in Figure 5. The authors fit a quadratic model to these data to explain the protein content given the yield. The fitted quadratic model and other fitted models are given in Table 4. The t value for the second-order coefficient in the quadratic model is 2.20, with a p value of .043. The cubic model does not provide any improvement. As the data show, point 4 might be regarded as an outlier. Table 4 shows the residual standard deviation for the different models fitted to both the complete data and the dataset when observation 4 is deleted.

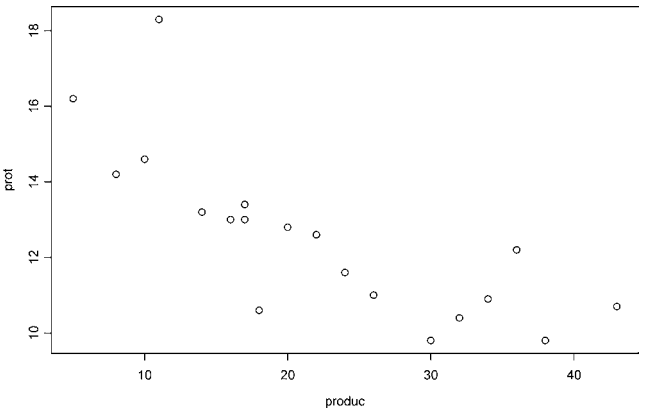

Figure 5. Graph of the Protein Data.

Table 4. Different Polynomial Models Fitted to the Protein Content Data

Order	*Model* $\widehat{y} =$	$\widehat{s}_R$	$\widehat{s}_{R(4)}$
0	12.54	2.212	1.767
1	$16.05 - \underset{(-5.16)}{.158x}$	1.420	1.012
2	$18.67 - \underset{(-3.37)}{.437x} + \underset{(2.20)}{.00587x^2}$	1.282	.8410
3	$17.98 - \underset{(-.81)}{.312x} - \underset{(-.015)}{.00027x^2} + \underset{(.35)}{.000087x^3}$	1.319	.8704
4	$14.43 + \underset{(-.55)}{.606x} - \underset{(-.88)}{.0073x^2} + \underset{(.92)}{.0023x^3} - \underset{(-.90)}{.000023x^4}$	1.328	.9031

NOTE: The third column shows the standard deviations of the residuals and the fourth column $\widehat{s}_{R(4)}$ is the standard deviation when point 4 is deleted.

Table 5 gives the posterior probabilities for each order using the three procedures and the three priors and the model selected by the AIC, BE, and FS procedures. In all cases IBF selects the linear model, except for $d = 4$ with the binomial and uniform prior. The BIC and FBF methods work similarly, selecting the linear model with the geometric prior and the quadratic model for the other two priors. The AIC criterion and BE for $d \geq 2$ choose the quadratic model, whereas FS chooses the linear model in all cases. Note that when the quadratic model has the highest probability, this value is relatively small, so that in effect, there is much uncertainty about the right model.

Table 6 gives the posterior probability for each order when data point 4 is deleted. All of the posterior probabilities for the quadratic model increase a bit and decrease for the linear model, except for the case of IBF with $d = 4$. BIC, FBF, AIC, and BE choose the same models as in Table 5, and now FS chooses the quadratic model instead of the linear model.

5.2 The Voltage Data

Montgomery and Peck (1992, p. 212) gave 41 observations on the battery voltage drop in a guided missile motor over time. A scatterplot of both variables is shown in Figure 6. These authors fitted a cubic spline with two knots to these data, obtaining a residual standard deviation of .2678. An alternative could be to fitted a polynomial regression model to these data. Montgomery and Peck stated that the cubic polynomial regression shows a cyclical pattern in the residuals. Figure 7 shows that this pattern disappears when fitting a polynomial of fourth degree. Table 7 gives the residual variance for several orders; it can be seen that the fourth-order model seems to fit the data quite well.

As in the previous example, Table 8 presents the results for the posterior probabilities of each model with degrees from 2 to 6 with $q = .15$ for the geometric prior. All methods choose the model of degree 4 (or the highest degree when the maximum degree d is <4). This example is interesting because it demonstrates good agreement of the IBF, BIC, and FBF methods in choosing a high-degree polynomial even in the case in which a prior penalizing the degree of the polynomial is selected. The AIC, FS, and BE methods also exhibit good behavior in all cases.

5.3 Growth Rate Data

The data comprise 10 samples with growth rate data for experimental rats fed by various doses of a dietary supplement, and come from Box et al. (1978, p. 480). These authors concluded that in view of the graphics and the analysis of variance table, the quadratic equation supplies an adequate representation over the region studied. A scatterplot of the data is given in Figure 8.

Table 9 shows that for $d = 1$, all methods choose the constant model and for $d \geq 2$, all methods choose the quadratic model except FS, which chooses the linear model in the first case and the cubic in the third. Note that with $d = 3$, the uncertainty about the best model can be important, and some methods give to the third-order-degree model a probability as high as .36, which will have a clear effect on the forecast generated by BMA.

6. CONCLUDING REMARKS

In this article we have carried out a comparative study of three methods to estimate the degree of a polynomial model and to obtain HDI for prediction. The three methods are compared with three different priors, two that penalize the degree of the polynomial and one that is uniform over the space of the model.

We conclude that IBF performs better than the other two methods, FBF and BIC, in selecting the model. Regarding the three priors used, the binomial seems to work better. For prediction purposes, whatever method is used, prediction intervals

Table 5. Posterior Probability of the jth-Order Model for the Protein Data

d, j	*IBFG*	*IBFB*	*IBFU*	*BICG*	*BICB*	*BICU*	*FBFG*	*FBFB*	*FBFU*	*AIC*	*BE*	*FS*
1, 0	.0393	.0020	.0031	.0074	.0004	.0006	.0336	.0017	.0026	0	**1**	0
1, 1	**.9607**	**.9980**	**.9969**	**.9926**	**.9996**	**.9994**	**.9664**	**.9983**	**.9974**	**1**	0	**1**
2, 0	.0352	.0018	.0026	.0061	.0001	.0001	.0309	.0008	.0011	0	0	0
2, 1	**.9631**	**.9799**	**.9754**	**.8195**	.2972	.2606	**.8696**	.4400	.3956	0	0	**1**
2, 2	.0017	.0183	.0219	.1744	**.7027**	**.7393**	.0995	**.5593**	**.6034**	**1**	**1**	0
3, 0	.0346	.0012	.0017	.0061	.0001	.0001	.0371	.0008	.0009	0	0	0
3, 1	**.9357**	**.7389**	**.6978**	**.8169**	.2734	.2203	**.8300**	.3387	.2790	0	0	**1**
3, 2	.0294	.2549	.2892	.1738	**.6464**	**.6249**	.1307	**.5925**	**.5856**	**1**	**1**	0
3, 3	.0001	.0050	.0112	.0032	.0800	.1547	.0023	.0680	.1345	0	0	0
4, 0	.0321	.0008	.0011	.0061	.0001	.0001	.0490	.0009	.0010	0	0	0
4, 1	**.8786**	.4666	.4169	**.8168**	.2709	.2077	**.8057**	.2995	.2242	0	0	**1**
4, 2	.0890	**.5259**	**.5666**	.1738	**.6404**	**.5893**	.1415	**.5844**	**.5250**	**1**	**1**	0
4, 3	.0002	.0052	.0108	.0032	.0793	.1459	.0038	.1050	.1886	0	0	0
4, 4	0	.0002	.0017	.0001	.0093	.0569	.0001	.0102	.0612	0	0	0

Table 6. Posterior Probability of the jth-Order Model for the Protein Data When the Point 4 Is Deleted

d, j	IBFG	IBFB	IBFU	BICG	BICB	BICU	FBFG	FBFB	FBFU	AIC	BE	FS
1, 0	.0104	.0005	.0008	.0014	.0001	.0001	.0094	.0005	.0007	0	**1**	0
1, 1	**.9896**	**.9995**	**.9992**	**.9986**	**.9999**	**.9999**	**.9906**	**.9995**	**.9993**	**1**	0	**1**
2, 0	.0108	.0004	.0005	.0008	0	0	.0080	.0001	.0001	0	0	0
2, 1	**.8957**	.4616	.4167	**.5301**	.0923	.0781	**.7394**	.2085	.1800	0	0	0
2, 2	.0929	**.5380**	**.5827**	.4691	**.9077**	**.9219**	.2526	**.7914**	**.8199**	**1**	**1**	**1**
3, 0	.0093	.0002	.0002	.0008	0	0	.0106	.0001	.0001	0	0	0
3, 1	**.8256**	.3100	.2718	**.5257**	.0834	.0641	**.6876**	.1577	.1241	0	0	0
3, 2	.1650	**.6888**	**.7260**	.4653	**.8198**	**.7570**	.2969	**.7567**	**.7143**	**1**	**1**	**1**
3, 3	.0001	.0006	.0013	.0082	.0969	.1789	.0050	.0855	.1614	0	0	0
4, 0	.0103	.0002	.0003	.0008	0	0	.0159	.0002	.0002	0	0	0
4, 1	**.9283**	**.5593**	**.4881**	**.5256**	.0828	.0615	**.6767**	.1460	.1054	0	0	0
4, 2	.0602	.4066	.4350	.4652	**.8142**	**.7263**	.2993	**.7175**	**.6217**	**1**	**1**	**1**
4, 3	.0008	.0309	.0646	.0082	.0962	.1716	.0080	.1273	.2207	0	0	0
4, 4	0	.0011	.0067	.0001	.0068	.0405	.0001	.0090	.0520	0	0	0

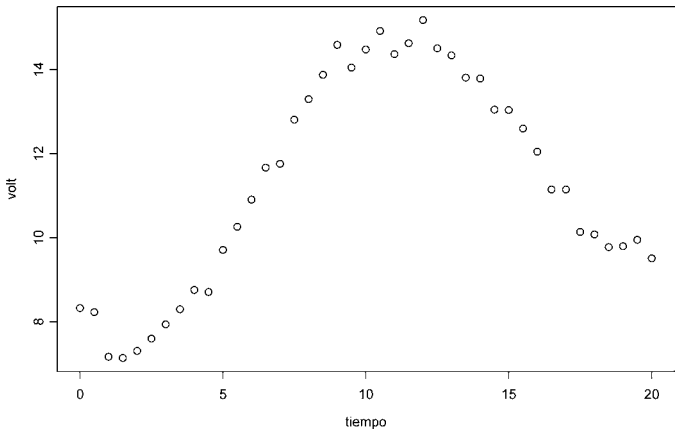

Figure 6. Graph of the Voltage Data.

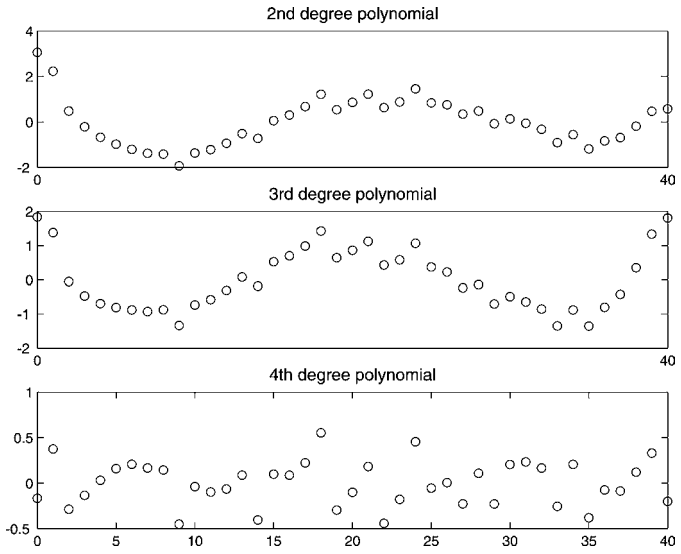

Figure 7. Residuals for Different Degrees of the Polynomial Model in the Voltage Data.

Table 7. Residual Standard Deviation for the Voltage Data for Different Polynomial Degrees

Order	0	1	2	3	4	5	6
Residual std	2.563	2.345	1.076	.9335	.2576	.2609	.2640

computed by BMA have a higher precision than those corresponding to the best model. These latter intervals are underestimated, and the BMA prediction appears to correct this effect.

It would be interesting to explore whether the results obtained in this article can be generalized to the case of several exploratory variables as in response surface methodology. This problem is now the subject of further research by us.

ACKNOWLEDGMENTS

The authors thank MCYT, Spain, for their support under grant BEC2000-0167.

APPENDIX: COMPARING PENALTY FUNCTIONS

Let $p_A(M_j|D)$ be the posterior probabilities of model M_j using method A, where $A = \{FBF; BIC\}$ assuming a uniform prior. For the FBF, the posterior probabilities are given by

$$p_{FBF}(M_j|D) = K\frac{\Gamma(\frac{n-j-1}{2})}{\Gamma(\frac{d-j+1}{2})}R_j^{-(n-d-2)/2}$$

whereas for the BIC approximation, the posterior probabilities are

$$p_{BIC}(M_j|D) = KR_j^{-n/2}n^{-(j+2)/2}.$$

These posterior probabilities have a similar functional form but differ in their penalty function, which is given by

$$pn_{BIC}(n,j) = n^{-(j+2)/2}$$

and

$$pn_{FBF}(n,j,d) = \frac{\Gamma(\frac{n-j-1}{2})}{\Gamma(\frac{d-j+1}{2})}.$$

We note that the penalty function for the BIC method, pn_{BIC}, is decreasing with n, whereas the penalty function for the FBF, pn_{FBF}, is increasing with $n^{n/2}$. To show this, using Stirling's approximation, we have

$$\log\Gamma(x+1) \approx \frac{1}{2}\log(2\pi) + \left(x+\frac{1}{2}\right)\log x - x,$$

Table 8. Posterior Probability of the jth-Order Model for the Voltage Data Where the Maximum Degree Is d

d, j	IBFG	IBFB	IBFU	BICG	BICB	BICU	FBFG	FBFB	FBFU	AIC	BE	FS
2, 0	0	0	0	0	0	0	0	0	0	0	0	0
2, 1	0	0	0	0	0	0	0	0	0	0	0	0
2, 2	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1**	**1**	**1**
3, 0	0	0	0	0	0	0	0	0	0	0	0	0
3, 1	0	0	0	0	0	0	0	0	0	0	0	0
3, 2	**.8097**	.3896	.2419	.1256	.0211	.0107	.2674	.0519	.0266	0	0	0
3, 3	.1903	**.6104**	**.7581**	**.8744**	**.9789**	**.9893**	**.7326**	**.9481**	**.9734**	**1**	**1**	**1**
4, 0	0	0	0	0	0	0	0	0	0	0	0	0
4, 1	0	0	0	0	0	0	0	0	0	0	0	0
4, 2	0	0	0	0	0	0	0	0	0	0	0	0
4, 3	0	0	0	0	0	0	0	0	0	0	0	0
4, 4	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1**	**1**	**1**
5, 0	0	0	0	0	0	0	0	0	0	0	0	0
5, 1	0	0	0	0	0	0	0	0	0	0	0	0
5, 2	0	0	0	0	0	0	0	0	0	0	0	0
5, 3	0	0	0	0	0	0	0	0	0	0	0	0
5, 4	**.9981**	**.9958**	**.9755**	**.9877**	**.9730**	**.8571**	**.9892**	**.9763**	**.8730**	**1**	**1**	**1**
5, 5	.0019	.0042	.0245	.0123	.0270	.1429	.0108	.0237	.1270	0	0	0
6, 0	0	0	0	0	0	0	0	0	0	0	0	0
6, 1	0	0	0	0	0	0	0	0	0	0	0	0
6, 2	0	0	0	0	0	0	0	0	0	0	0	0
6, 3	0	0	0	0	0	0	0	0	0	0	0	0
6, 4	**.9994**	**.9987**	**.9920**	**.9875**	**.9726**	**.8364**	**.9830**	**.9630**	**.7918**	**1**	**1**	**1**
6, 5	.0006	.0013	.0080	.0123	.0270	.1394	.0168	.0366	.1807	0	0	0
6, 6	0	0	0	.0002	.0003	.0242	.0002	.0004	.0275	0	0	0

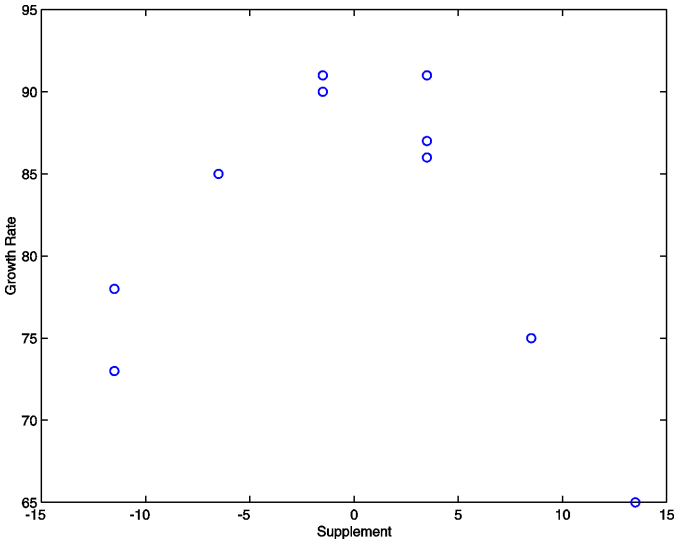

Figure 8. Graph of the Growth Rate Data.

so that

$$\log\bigl(pn_{FBF}(n,j,d)\bigr) = \log\Gamma\left(\frac{n-j-1}{2}\right) - \log\Gamma\left(\frac{d-j+1}{2}\right),$$

$$\begin{aligned}\log\bigl(pn_{FBF}(n,j,d)\bigr) \approx{}& \frac{1}{2}(n-j-2)\log(n-j-3)\\ &- \frac{1}{2}(d-j-2)\log(d-j-3)\\ &+ \frac{1}{2}(-n+d+6)\log 2,\end{aligned}$$

and

$$\begin{aligned}&\log\bigl(pn_{FBF}(n,j,d)\bigr)\\ &\quad\approx \frac{1}{2}(n-j-2)\log(n-j-3) - \frac{n}{2}\log 2 + h(j,d).\end{aligned}$$

To compare these penalty functions, we standardize them to sum to 1, yielding

$$pns_{BIC}(j) = \frac{pn_{BIC}(j)}{\sum_{j=0}^{d} pn_{BIC}(j)}$$

Table 9. Posterior Probability of the jth-Order Model for the Growth Data

d, j	IBFG	IBFB	IBFU	BICG	BICB	BICU	FBFG	FBFB	FBFU	AIC	BE	FS
1, 0	**.9598**	**.5443**	**.6418**	**.9725**	**.6389**	**.7263**	**.9742**	**.6536**	**.7389**	**1**	**1**	0
1, 1	.0402	.4557	.3582	.0275	.3611	.2737	.0258	.3464	.2611	0	0	**1**
2, 0	.2636	.0016	.0020	.0018	0	0	.2132	.0012	.0015	0	0	0
2, 1	.0068	.0007	.0006	.0001	0	0	.0087	.0010	.0008	0	0	0
2, 2	**.7289**	**.9976**	**.9974**	**.9981**	**1.0000**	**1.0000**	**.7780**	**.9978**	**.9976**	**1**	**1**	**1**
3, 0	.2340	.0013	.0016	.0018	.0000	.0000	.3388	.0020	.0021	0	0	0
3, 1	.0144	.0019	.0014	.0001	0	0	.0173	.0020	.0014	0	0	0
3, 2	**.7250**	**.7748**	**.6338**	**.9669**	**.8227**	**.6988**	**.6234**	**.8171**	**.6930**	**1**	**1**	0
3, 3	.0143	.2208	.3614	.0313	.1773	.3012	.0205	.1789	.3034	0	0	**1**

and

$$pns_{FBF}(j) = \frac{pn_{FBF}(j)}{\sum_{j=0}^{d} pn_{FBF}(j)}$$

so that, after some algebra, we have

$$p_{BIC}(M_j|D) = (KL_7)(pns_{BIC}(j))R_j^{-n/2}$$

and

$$p_{FBF}(M_j|D) = (KL_5)(pns_{FBF}(j))_j^{-(n-d-2)/2},$$

where $L_7 = \sum_{j=0}^{d} pn_{BIC}(j)$ and $L_5 = \sum_{j=0}^{d} pn_{FBF}(j)$, so that the standardized penalty constants are grouped with the standardized constants.

[Received May 2002. Revised June 2004.]

REFERENCES

Abramowitz, M., and Stegun, I. (1970), *Handbook of Mathematical Functions*, New York: Dover Publications, Inc.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings of the 2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Czaki, Budapest, Hungarian: Akademiai Kiado, pp. 267–281.

Berger, J., and Pericchi, L. (1996a), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics V*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 25–44.

——— (1996b), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Box, G., Hunter, W., and Hunter, J. (1978), *Statistics for Experimenters*, New York: Wiley.

Chatfield, C. (1995), "Model Uncertainty, Data Mining and Statistical Inference" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 158, 419–466.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–24.

Draper, N., and Guttman, I. (1987), "A Common Model Selection Criterion," in *Proceeding of the Symposium on Probability and Bayesian Statistics*, New York: Plenum Publisher Corporation, pp. 139–150.

Fernández, C., Ley, E., and Steel, M. (2002), "Bayesian Modeling of Catch in a Northwest Atlantic Fishery," *Journal of the Royal Statistical Society*, Ser. A, 51, 257–280.

George, E. (1999), *Bayesian Model Selection*, New York: Wiley.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417.

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

Leamer, E. (1978), *Bayesian Statistics: An Introduction*, New York: Wiley.

Montgomery, D., and Peck, E. (1992), *Introduction to Linear Regression Analysis*, New York: Wiley.

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

Philips, R., and Guttman, I. (1998), "A New Criterion for Variable Selection," *Statistics and Probability Letters*, 38, 11–19.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Model," *Journal of the American Statistical Association*, 92, 179–191.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Snedecor, G., and Cochran, W. (1989), *Statistical Methods* (8th ed.), Ames, IA: Iowa State University Press.

Thompson, M. (1978), "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation," *International Statistical Review*, 46, 1–19.