UNIVERSIDAD CARLOS III DE MADRID

working papers

Working Paper 03-21
Statistics and Econometrics Series 04
April 2003

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

A Bayesian Approach for Predicting with Polynomial Regresión of Unknown Degree.

Irwin Guttman, Daniel Peña and M Dolores Redondas*

Abstract

This article presents a comparison of four methods to compute the posterior probabilities of the possible orders in polynomial regression models. These posterior probabilities are used for forecasting by 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 are illustrated with simulations and some known engineering data.

Key words:

Bayesian Model Averaging; Fractional Bayes Factor; Intrinsic Bayes Factor; Bayesian Information Criterion.

*Guttman, State University of New York , Buffalo. U.S.A. Peña, Statistics and Econometrics Department, University Carlos III of Madrid, Spain, e-mail: dpena@est-econ.uc3m.es, Redondas, Statistics and Econometrics Department, University Carlos III of Madrid, Spain, e-mail: redondas@est-econ.uc3m.es.Tel: +34 916249314. We also acknowledge financial support from BEC2000-0167, MCYT, Spain.

A Bayesian Approach for Predicting with Polynomial Regression of Unknown Degree

Irwin Guttman, Daniel Peña and Dolores Redondas

April 3, 2003

Abstract

This article presents a comparison of four methods to compute the posterior probabilities of the possible orders in polynomial regression models. These posterior probabilities are used for forecasting by 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 are illustrated with simulations and some known engineering data.

Key words: Bayesian Model Averaging; Fractional Bayes factor; Intrinsic Bayes factor; Bayesian Information Criterium.

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 point of view, 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 criteria, such as that of Akaike (1973) and others. From the Bayesian point of view 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 Schwarz (1978) criterion, 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 paper we first compare different procedures for computing the posterior probabilities for different polynomial degrees and then we take into account the model uncertainty for forecasting using Bayesian model averaging (BMA).

The problem of computing posterior probabilities for models using non informative priors has been the subject of much recent research. Several authors have proposed the use of training samples in order to avoid the problem of the constant indetermination of the Bayes factors when improper priors are used, see Atkinson (1978), Gelfand and Chang (1992), Lempers (1971) and San Martini and Spezzaferri (1984). Spiegelhalter and Smith (1982) proposed the use of an imaginary training sample to determine the ratio of constants that appear in the Bayes factors. They assume a minimal training sample, which is such that this imaginary sample does not provide evidence in favor of any model. Geisser and Eddy (1979) base their analysis on part of the sample chosen in certain optimal ways. In a similar way, Gelfand and Ghosh (1998) proceed by minimizing certain posterior losses found for given models. See Gelfand and Dey (1994) for a review of the literature. O'Hagan (1995) proposed the use of fractional Bayes factors and shows that his method preserves the asymptotic properties of the Bayes factors. Finally, Berger and Pericchi (1996a, 1996b) proposed the use of the intrinsic Bayes factor. They used the training sample of minimal size which makes the posterior density of the parameter proper.

Bayesian model averaging leads to forecasts which are a weighted average of the predictive densities obtained by considering all possible polynomial degrees with weights equal to the posterior probabilities of each degree. See Draper and Guttman (1987). Accordingly, BMA takes into account the uncertainty about the different models, as was pointed out in the seminal work of Leamer (1978). This method was not operational in the past due to its heavy computational requirements, but the possibilities opened up by using MCMC methods have made the method computationally feasible. George (1999) reviews Bayesian model selection and discusses BMA in the context of decision theory, Draper (1995), Chatfield (1995) and Kass and Raftery (1995) review BMA and the cost of ignoring the model uncertainty. Hoeting et al. (1999) present a review of BMA emphasizing the implementation and practical matters. For linear regression models there is an extensive literature, see e.g. Raftery et al. (1997) who proposed the BMA implementation in MC^3 and Occam's window, Fernandez et al. (2002) who carry out BMA for a large number of possible regressors and also provide an automatic prior structure that can be used in these cases.

Some framework for the problem is as follows: Let $\{M_1, M_2, \ldots, M_K\}$ be the space of all the models under consideration and let $\mathbf{y}$ be the vector of observations. Suppose that the prior $p(\boldsymbol{\theta}_i|M_i) = c_i g(\boldsymbol{\theta}_i)$ is improper, that is 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

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

where $m\left(\mathbf{y}\right) = \sum_{i=1}^{K} p\left(\mathbf{y}\left|M_i\right.\right)p\left(M_i\right)$ and $p(M_i)$ is the prior probability that M_i holds. This probability depends on the unknown constant c_i. Thus, we have that the Bayes factor for comparing two models, M_i and M_j, is

$$B_{ij} = \frac{p\left(M_i|\mathbf{y}\right)}{p\left(M_j|\mathbf{y}\right)} = \frac{c_i}{c_j}\frac{\left[\int p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right)g(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i\right]}{\left[\int p\left(\mathbf{y}|\boldsymbol{\theta}_j, M_j\right)g(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j\right]}\frac{p(M_i)}{p(M_j)} \tag{2}$$

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

For a given model M_i the posterior predictive distribution, $p(y_f|\mathbf{y}, M_i)$ when predicting a future observation, y_f, is given by

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

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

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

We will use (4) in the sequel and refer to it is Bayesian Model Averaging (or BMA for short), for indeed (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 as, inserting (3) in (4),

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

which takes into account both the parameter variability and the model variability.

This paper is organized as follows. In Section 2, the polynomial model is presented and an informative prior for the model space is introduced. This prior favors the parsimony principle with respect to the degree of the polynomial. In Section 3, we present four different approaches to the problem. The first one is a procedure for computing the posterior probabilities of the model based on the work by Philips and Guttman (1998). This method averages over subsets of all possible available training samples to avoid the possible sensitivity of a particular training sample in this problem. The second one is the intrinsic Bayes factor of Berger and Pericchi (1996b). The third one is an approximate method based on the Bayesian Information Criterion (BIC), proposed by Schwarz (1978) and the fourth is the fractional Bayes factor proposed by O'Hagan (1995). The last subsection presents a brief summary of these four methods. In Section 4, we study the predictive distribution for this problem. The methods presented are compared in a Monte Carlo study in Section 5, and using some real data examples in Section 6. Finally, Section 7 gives some concluding remarks.

2 The polynomial model and the prior.

We focus here on the general polynomial regression model, M_j

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

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

$$\mathbf{y} = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon} \tag{5}$$

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

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

To compute the posterior probabilities that model M_j holds (i.e., the degree is j) using (1), we would need the normalizing constant $(m(\mathbf{y}))^{-1}$. However, the normalizing constant cannot at this point be determined simply by using the fact that $\sum_{j=1}^{K} p(M_j|\mathbf{y}) = 1$, because the models under consideration have different dimensional parameter space (we discuss this problem in Section 3 of this paper). But we do note that we may find $p(M_j|\mathbf{y})$, say, by employing Bayes factors B_{ij}, for it is straightforward to show that

$$p(M_j|D) = \frac{p(D|M_j)\, p(M_j)}{\sum_{i=1}^{K} p(D|M_i)\, p(M_i)} = \left[\sum_{i=1}^{K} B_{ij} \frac{p(M_i)}{p(M_j)}\right]^{-1}, \tag{6}$$

where, $B_{ij} = p(D|M_i)/p(D|M_j)$ is the needed Bayes factor. It is important to note that when improper priors are used, the Bayes factors do depend on the unknown indeterminate ratio c_i/c_j -see (2). Also, we have used the notation D to denote the data $(\mathbf{X}_d, \mathbf{y})$ with the understanding that for Model i, $i < k$, then a subset of D is to be used, namely $(\mathbf{X}_i, \mathbf{y})$.

2.1 The prior for the models

We consider two 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

$$p(M_j) = (d+1)^{-1}. \tag{7}$$

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,

$$p(M_j) = \frac{(1-q)}{(1-q^{d+1})} q^j \quad j = 0, 1, 2, \ldots, d \tag{8}$$

for $0 < q < 1$, where j is the degree of the model. We are interested in choosing a model, given the data, as parsimonious as possible, and, with this aim, we have chosen the prior (8) that favors M_0, so that a priori $E(\mathbf{Y}) = \beta_0$. Making a correspondence between J and M_j, this implies that we should choose the prior in such a way that $E(J) < 0.5$, that is, $E(J) = \frac{q}{(1-q)} \frac{1-(d+1)q^d+dq^{d+1}}{(1-q^{d+1})} < 0.5$, which as may be verified, holds if we choose $q < 1/3$. The prior (8) decreases as j increases, and has the advantage that the ratios $p(M_j)/p(M_{j+1})$ are constant for $j = 0, ..., d-1$.

3 Methods for obtaining the posterior probability of the models

3.1 The constant dimension method

As indicated in the previous section when using a reference prior for the parameters, we have the problem that different models M_j have parameter spaces of different dimensions. To avoid this problem, following Philips and Guttman (1998), we first redefine M_j as follows:

$$M_j : y = (1, x, ..., x^j, 0, ..., 0)\boldsymbol{\beta}_d + \epsilon,$$

where, as before, ϵ is $N(0, \sigma^2)$. Secondly, we select at random a training sample of size $m = d+2$ out of n observations. We index the use of a particular training sample of size m by t, $t = 1, ..., T = \begin{pmatrix} n \\ m \end{pmatrix}$. Then 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 $\mathbf{X}_d$ matrix, say $\mathbf{X}_t(d)$, pertain to the training sample, so that $D_t = (\mathbf{X}_t(d), \mathbf{y}_t)$ corresponds to the training sample. Suppose that the standard non informative prior for $(\boldsymbol{\beta}_d, \sigma^2)$ is used. Then, the posterior distribution for the parameters given the training sample is

$$p(\boldsymbol{\beta}_d, \sigma^2|D_t) = K_1(\sigma^2)^{-(\frac{m}{2}+1)} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}_t - \mathbf{X}_t(d)\boldsymbol{\beta}_d)'(\mathbf{y}_t - \mathbf{X}_t(d)\boldsymbol{\beta}_d)\right) \tag{9}$$

where K_1 is a constant depending only on t. Now, we use the posterior (9) as a prior for the remaining analysis. We need the following notation. Let the $(n-m) \times (d+1)$ matrix $\mathbf{X}_{(-t)}(d)$ and the $(n-m)\times 1$ vector $\mathbf{y}_{(-t)}$ be the corresponding matrix and vector obtained by deleting the m rows from $\mathbf{X}_d$ and $\mathbf{y}$. Further, we denote by $\mathbf{X}_{(-t)}\,(j)$ a $(n-m)\times(d+1)$ matrix obtained from $\mathbf{X}_{(-t)}(d)$ that has its last $d-j$ columns replaced by $(n-m)\times 1$ vectors of zeros. We may now denote a new $n \times (d+1)$ matrix appropriate to the assumption that M_j holds by

$$Z_j = \left(\begin{array}{c} \mathbf{X}_t(d) \\ \mathbf{X}_{(-t)}\,(j) \end{array} \right) \tag{10}$$

and the partitioned vector of observations by

$$\mathbf{z} = \left(\begin{array}{c} \mathbf{y}_t \\ \mathbf{y}_{(-t)} \end{array} \right),$$

to be used, assuming M_j holds, in the model

$$E\,(\mathbf{z}) = \mathbf{Z}_j\boldsymbol{\beta}_d.$$

The joint posterior for the parameters and a particular model is

$$p(\boldsymbol{\beta}_d, \sigma^2, M_j|D, t) \propto p(D|M_j, \boldsymbol{\beta}_d, \sigma^2)p(\boldsymbol{\beta}_d, \sigma^2|D_t)p(M_j)$$

where we have assumed independence of M_j and $(\boldsymbol{\beta}_d, \sigma^2)$ a priori, and the prior $p(\boldsymbol{\beta}_d, \sigma^2|D_t)$ is given by (9). Thus, we can compute the joint posterior by combining this prior with the likelihood arising from the polynomial regression model based on $\mathbf{y}_{(-t)}$ and $\mathbf{X}_{(-t)}\,(j)$ and we obtain

$$p(\boldsymbol{\beta}_d, \sigma^2, M_j|D, t) \propto p(M_j)(\sigma^2)^{-(n/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{z} - \mathbf{Z}_j\boldsymbol{\beta}_d)'(\mathbf{z} - \mathbf{Z}_j\boldsymbol{\beta}_d)\right). \tag{11}$$

The result (11) arises from the assumption $E(\mathbf{y}_{(-t)}|M_j) = \mathbf{X}_{(-t)}\,(j)\,\boldsymbol{\beta}_d$ which in turn implies $\beta_{j+1} = ... = \beta_d = 0$ so that M_j is a jth degree polynomial. We note that $\mathbf{Z}_j$ has non zero entries in the first m rows of its first $d-j$ columns, so that under the plausible assumption that this matrix is non singular, we can write

$$\begin{aligned} p(\boldsymbol{\beta}_d, \sigma^2, M_j|D, t) &\propto p(M_j)(\sigma^2)^{-(n/2+1)} \\ &\exp\left[-\frac{1}{2\sigma^2}\left(S_j + (\boldsymbol{\beta}_d - \widehat{\boldsymbol{\beta}}_{dj})'\mathbf{Z}_j'\mathbf{Z}_j(\boldsymbol{\beta}_d - \widehat{\boldsymbol{\beta}}_{dj})\right)\right] \end{aligned} \tag{12}$$

where

$$\widehat{\boldsymbol{\beta}}_{dj} = (\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\mathbf{z}$$

and the sum of squares of the residuals is given by

$$S_j = \mathbf{z}'(\mathbf{I}_n - \mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j')\mathbf{z}. \tag{13}$$

Finally, the model posterior probabilities are found integrating (12) over the $d+2$ parameter space, and we obtain

$$p(M_j|D,t) = K_2(S_j^{n-d-1}\left|\mathbf{Z}_j'\mathbf{Z}_j\right|)^{-1/2}p(M_j)$$

where K_2 is given by

$$K_2 = \left[\sum_{j=0}^{d}\left(S_j^{n-d-1}\left|\mathbf{Z}_j'\mathbf{Z}_j\right|\right)^{-1/2}p(M_j)\right]^{-1}$$

The posterior probabilities depend on the training sample and, if the sample size is not large, this dependency could be important. In order to avoid this problem we may compute $p(M_j|D)$ by averaging over all $T = \left(\begin{array}{c} m \\ n \end{array}\right)$ possible training samples of size m. Then, we define

$$\begin{aligned} p(M_j|D) &= \frac{1}{T}\sum_{t=1}^{T} p(M_j|D,t) = \\ &= \frac{1}{T}\sum_{t=1}^{T} K_2(S_j^{n-d-1}\left|\mathbf{Z}_j'\mathbf{Z}_j\right|)^{-1/2}p(M_j) \end{aligned}$$

Now, we recall that the prior (8) penalizes the degree of the polynomials and if this prior is employed, we readily find that

$$p(M_j|D) = K_3\frac{1}{T}\sum_{t=1}^{T} q^j(S_j^{n-d-1}\left|\mathbf{Z}_j'\mathbf{Z}_j\right|)^{-1/2}.$$

The computation of this probability may involve a large number of terms. Our proposal then is to take two subsets of training samples of size T_0 and based on each compute $p(M_j|D,T_0^{(1)})$ and $p(M_j|D,T_0^{(2)})$. If

$$\sup_j \left|p(M_j|D,T_0^{(1)}) - p(M_j|D,T_0^{(2)})\right| \leq \delta \tag{14}$$

for some small δ we stop. Otherwise, we increase the size of T_0 until the previous condition is met.

3.2 Intrinsic Bayes factors.

Berger and Pericchi (1996a, 1996b) proposed to solve the indetermination problem by using intrinsic Bayes factors. The idea of the procedure is as follows. Note that

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

where $B_{ij}\left(t\right)$ 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

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

Here, $D_{(-t)}$ refers to the data $\left(\mathbf{X}_{(-t)}, \mathbf{y}_{(-t)}\right)$.

Suppose that we use non informative priors, so that B_{ij} and B_{ji}^{t} depend, as shown in (2), on unknown constants. Then, these constants will be cancelled out when computing the conditional Bayes factor. As the conditional Bayes factor depends on the training sample, Berger and Pericchi propose several types of averaging over all the possible training samples. One of their proposals is the use of the arithmetic intrinsic Bayes factor which is defined as follows

$$B_{ij}^{AI} = \frac{1}{T}\sum_{i=1}^{T} B_{ij}(t).$$

For normal errors, Berger and Pericchi (1996a) find that $B_{ij}(t)$ can be computed by

$$B_{ij}(t) = C_{ij}\frac{\left|\mathbf{X}_j'\mathbf{X}_j\right|^{1/2}\left|\mathbf{X}_t'(i)\mathbf{X}_t(i)\right|^{1/2}}{\left|\mathbf{X}_i'\mathbf{X}_i\right|^{1/2}\left|\mathbf{X}_t'(j)\mathbf{X}_t(j)\right|^{1/2}}\frac{R_j^{(n-j-1)/2}R_t(i)^{1/2}}{R_i^{(n-i-1)/2}R_t(j)^{1/2}},$$

where $\mathbf{X}_j$ is the $n\times(j+1)$ design matrix with the complete data, with columns $(1, x, ..., x^j)$, $\mathbf{X}_t(j)$ is the $t\times(j+1)$ formed by the rows of $\mathbf{X}_j$ corresponding to the training sample, $R_j = \mathbf{y}'(\mathbf{I}_n - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j')\mathbf{y}$ is the sum of square of the residuals for $\mathbf{X}_j$, $R_t\left(i\right) = \mathbf{y}_{(t)}'(\mathbf{I}_t - \mathbf{X}_t(j)(\mathbf{X}_t'(j)\mathbf{X}_t(j))^{-1}\mathbf{X}_t'(j))\mathbf{y}_{(t)}$ is the sum of square of the residuals for the training sample, and

$$C_{ij} = \frac{\Gamma\left(\frac{n-i-1}{2}\right)\Gamma\left(\frac{i-j+1}{2}\right)}{\Gamma\left(\frac{n-j-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}.$$

Then, the posterior probabilities can be computed by

$$p\left(M_j|D\right) = \left[\sum_{i=1}^{K} B_{ij}^{AI}\frac{p\left(M_i\right)}{p\left(M_j\right)}\right]^{-1}. \qquad (15)$$

The arithmetic intrinsic Bayes factor is very expensive to compute, and as Berger and Pericchi commented, is unstable for small sample sizes. With this as background, the authors recommended the use of the expected intrinsic Bayes factor, which in nested models with $M_i \subset M_j$, can be computed by

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

where

$$C_{ij}^{*} = C_{ij} \left(\Gamma\left(\frac{1}{2}\right)^{-1}\right) \left(\frac{n-i-1}{2}\right)^{(j-i)/2}, \tag{16}$$

$$\lambda_{ij}\left(t\right) = \frac{R_i}{n-j-1} \boldsymbol{\beta}_i' \mathbf{X}_i'\left(l\right) \left[\mathbf{I} - \mathbf{X}_i\left(t\right) \left(\mathbf{X}_i'\left(t\right) \mathbf{X}_i\left(t\right)\right)^{-1} \mathbf{X}_i'\left(t\right)\right]^{-1} \mathbf{X}_i\left(l\right) \boldsymbol{\beta}_i, \tag{17}$$

and $M\left(a,b,c\right)$ is Kummer's function (see Abramowitz and Stegun, 1970, chapter 13). Then, they define $B_{ji}^{EAI} = 1/B_{ij}^{EAI}$. We use the expected intrinsic Bayes factor for the comparison of the posterior probabilities.

3.3 Fractional Bayes factor

O'Hagan (1995) proposed the use of a modified Bayes factor to avoid the problem of indetermination when non informative priors are used. This is called fractional Bayes factor. For a data set $D = \left(\mathbf{y}, \mathbf{X}\right)$, it is defined as

$$B_{ij}^{b}(D) = \frac{q_i\left(b, D\right)}{q_j\left(b, D\right)}, \tag{18}$$

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

$$q_i\left(b, D\right) = \frac{\int g\left(\boldsymbol{\theta}_i\right) p_i\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) d\theta_i}{\int g\left(\boldsymbol{\theta}_i\right) \left[p_i\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right)\right]^b d\theta_i}, \tag{19}$$

$g\left(\boldsymbol{\theta}_i\right)$ the prior distribution for the parameters, and where $p_i\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right)$ is the full likelihood, under the model M_i. Note that if $b = 0$, that is there is no training sample involved so that (18) 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

$$p\left(M_j|D\right) = \left[\sum_{i=1}^{K} B_{ij}^{b} \frac{p\left(M_i\right)}{p\left(M_j\right)}\right]^{-1}. \tag{20}$$

We may now compute $q_j\left(b, D\right)$ for the polynomial model (5) using a non informative prior for the parameters $\boldsymbol{\theta}_j = \left(\boldsymbol{\beta}_d, \sigma^2\right)$, given by $g\left(\boldsymbol{\theta}_j\right) = p\left(\boldsymbol{\beta}_d, \sigma^2\right) \propto \sigma^{-2}$. Then, the denominator of $q_i\left(b, D\right)$ in (19) is

$$\int g(\boldsymbol{\theta}_i)\left[p_i(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\right]^b d\theta_i = (2\pi)^{-nb/2} \times$$
$$\times \int \sigma^{-\left(\frac{nb}{2}+1\right)} \exp\left\{-\frac{b}{2\sigma^2}\left(R_i + (\boldsymbol{\beta}_d - \widehat{\boldsymbol{\beta}}_{di})'\mathbf{X}_i'\mathbf{X}_i(\boldsymbol{\beta}_d - \widehat{\boldsymbol{\beta}}_{di})\right)\right\} d\boldsymbol{\beta}_d d\sigma^2,$$

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

$$R_i = \mathbf{y}'(\mathbf{I}_n - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i')\mathbf{y}. \qquad (21)$$

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

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

where $w = nb - d - 1$ are the degrees of freedom. As the numerator of $q_i(b, D)$ is identical with $b = 1$, we then have that

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

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

$$B_{ki}^b(D) = \frac{\Gamma\left(\frac{n-k-1}{2}\right)\Gamma\left(\frac{nb-i-1}{2}\right)}{\Gamma\left(\frac{nb-k-1}{2}\right)\Gamma\left(\frac{n-i-1}{2}\right)}\left(\frac{R_k}{R_i}\right)^{-n(1-b)/2}, \qquad (22)$$

and from (20) the model posterior probabilities are

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

where

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

3.4 The BIC approximation

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

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

where $\widehat{\boldsymbol{\theta}}_i$ is the MLE of the parameter vector under model M_i and d_i is the dimension of the vector $\boldsymbol{\theta}_i$. The Bayesian information criterion (BIC) of a model M_i is

$$BIC\left(M_i\right) = -2S\left(M_i\right),$$

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

$$B_{ij}^{BIC} = \exp\left(S\left(M_i\right) - S\left(M_j\right)\right) = \frac{\exp\left(-0.5BIC\left(M_i\right)\right)}{\exp\left(-0.5BIC\left(M_j\right)\right)}$$

and obtain the posterior probability for a model by

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

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

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

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

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

where

$$K_{BIC} = \sum_{j=0}^{d} R_j^{-n/2} n^{-(j+2)/2}.$$

3.5 Some comparison of the methods

Let $p_{AB}\left(M_j\left|D\right.\right)$ be the posterior probabilities of model M_j using method A, where $A = \{CDM; IBF; FBF; BIC\}$, denotes the four methods presented in the previous sections, (constant dimension (CD), Intrinsic Bayes Factor (IBF), Fractional Bayes Factors (FBF) and the BIC approximation), and $B = \{I,\ NI\}$ denotes the use of informative or non informative priors for the models. Then, for the constant dimension method presented in section 3.1 the posterior probabilities are given by

$$
\begin{aligned}
p_{CDMI}(M_j|D) &= K_2 \frac{1}{T} \sum_{t=1}^{T} q^j (S_j^{n-d-1} \left|\mathbf{Z}_j^{'} \mathbf{Z}_j\right|)^{-1/2} \\
p_{CDMNI}(M_j|D) &= K_3 \frac{1}{T} \sum_{t=1}^{T} (S_j^{n-d-1} \left|\mathbf{Z}_j^{'} \mathbf{Z}_j\right|)^{-1/2}
\end{aligned}
$$

where $\mathbf{Z}_j$ and S_j are given by (10) and (13).

For the intrinsic Bayes factors presented in 3.2, we will use the expected intrinsic Bayes factor, which leads to

$$
\begin{aligned}
B_{ij}^{EAI} &= C_{ij}^{*} \frac{\left|\mathbf{X}_j^{\prime}\mathbf{X}_j\right|^{1/2} \left|\mathbf{X}_t^{\prime}(i)\mathbf{X}_t(i)\right|^{1/2}}{\left|\mathbf{X}_i^{\prime}\mathbf{X}_i\right|^{1/2} \left|\mathbf{X}_t^{\prime}(j)\mathbf{X}_t(j)\right|^{1/2}} \left(\frac{R_i}{R_j}\right)^{\frac{(n-i-1)}{2}} \\
&\times \exp\left(-\lambda_{ij}\left(t\right)/2\right) M\left(1/2, (j-i+1)/2, \lambda_{ij}\left(t\right)/2\right),
\end{aligned}
$$

where the constants C_{ij}^{*} and λ_{ij} are given by (16) and (17) respectively, with R_i defined by (21), so that the posterior probability can be obtained by

$$
\begin{aligned}
p_{IBFI}(M_i|D) &= \left(\sum_{k=1}^{K} \frac{q^j}{q^i} B_{ki}^{EAI}\right)^{-1} \\
p_{IBFNI}(M_i|D) &= \left(\sum_{k=1}^{K} B_{ki}^{EAI}\right)^{-1}
\end{aligned}
$$

For the fractional Bayes factor, presented in section 3.3, the posterior probabilities are given by

$$
\begin{aligned}
p_{FBFI}(M_j|D) &= K_4 q^j \frac{\Gamma\left(\frac{n-j-1}{2}\right)}{\Gamma\left(\frac{d-j+1}{2}\right)} R_j^{-(n-d-2)/2} \\
p_{FBFNI}(M_j|D) &= K_5 \frac{\Gamma\left(\frac{n-j-1}{2}\right)}{\Gamma\left(\frac{d-j+1}{2}\right)} R_j^{-(n-d-2)/2}
\end{aligned}
$$

Finally, the BIC approximations of the posterior probabilities are

$$
\begin{aligned}
p_{BICI}(M_j|D) &= K_6 q^j R_j^{-n/2} n^{-(j+2)/2} \\
p_{BICNI}(M_j|D) &= K_7 R_j^{-n/2} n^{-(j+2)/2}.
\end{aligned}
$$

The posterior probabilities computed by the FBF and BIC methods have a similar functional form but differ in the penalty function that is given by

$$pn_B\left(n,j\right) = n^{-(j+2)/2}$$
$$pn_F\left(n,j,d\right) = \frac{\Gamma\left(\frac{n-j-1}{2}\right)}{\Gamma\left(\frac{d-j+1}{2}\right)}$$

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

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

so that,

$$\log\left(pn_F\left(n,j,d\right)\right) = \log\Gamma\left(\frac{n-j-1}{2}\right) - \log\Gamma\left(\frac{d-j+1}{2}\right)$$
$$\log\left(pn_F\left(n,j,d\right)\right) \approx \frac{1}{2}\left(n-j-2\right)\log\left(n-j-3\right) - \frac{1}{2}\left(d-j-2\right)\log\left(d-j-3\right) + \frac{1}{2}\left(-n+d+6\right)\log 2$$
$$\log\left(pn_F\left(n,j,d\right)\right) \approx \frac{1}{2}\left(n-j-2\right)\log\left(n-j-3\right) - \frac{n}{2}\log 2 + h\left(j,d\right)$$

In order to compare these penalty functions we standardize them to sum to one, yielding

$$pns_B\left(j\right) = \frac{pn_B\left(j\right)}{\sum_j pn_B\left(j\right)}$$
$$pns_F\left(j\right) = \frac{pn_F\left(j\right)}{\sum_j pn_F\left(j\right)}$$

so that after some algebra, we have

$$p_{BICNI}(M_j|D) = \left(K_7L_7\right)\left(pns_B\left(j\right)\right)R_j^{-n/2}$$
$$p_{FBFNI}(M_j|D) = \left(K_5L_5\right)\left(pns_F\left(j\right)\right)R_j^{-(n-d-2)/2}$$

where $L_7 = \sum_j pn_B\left(j\right)$ and $L_5 = \sum_j pn_F\left(j\right)$ so that the standardized penalty constants are grouped with the standardized constants. The standardized penalty function of BIC depends on the maximum degree d, only through its denominator. Figure (1) shows these standardized penalty functions as a function of the sample size, n, for $j = 0, 1, 2$ with maximum degree $d = 5$, while Figure (2) does the same for $d = 10$. In both cases n is allowed to vary between 50 and 500. These figures show that BIC penalizes more than FBF, while BIC gives more weight to the model with lowest degree and thus it gives less weight than the FBF to polynomials of higher degree.

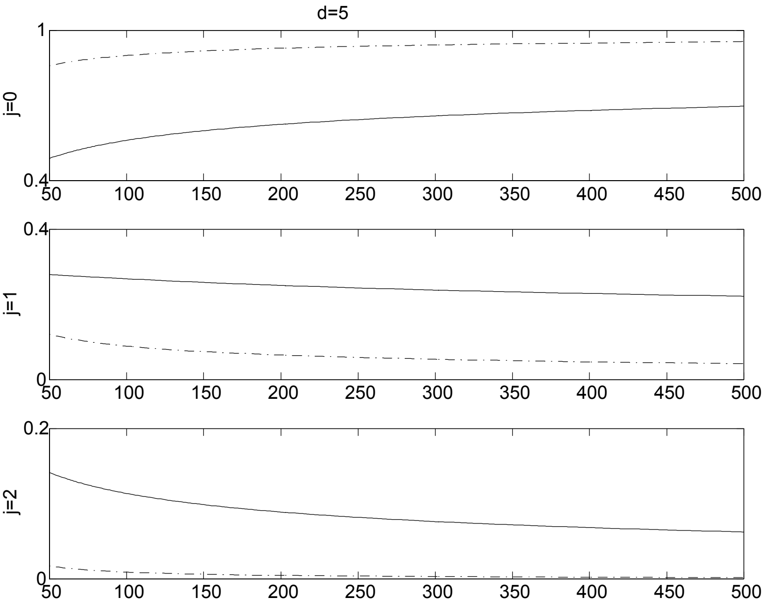

Figure 1: Standardized penalty of BIC method - dash-dot line-, and FBF method -solid line- where the maximum degree d is 5, for the models, constant, $j = 0$; linear, $j = 1$; and quadratic $j = 2$.

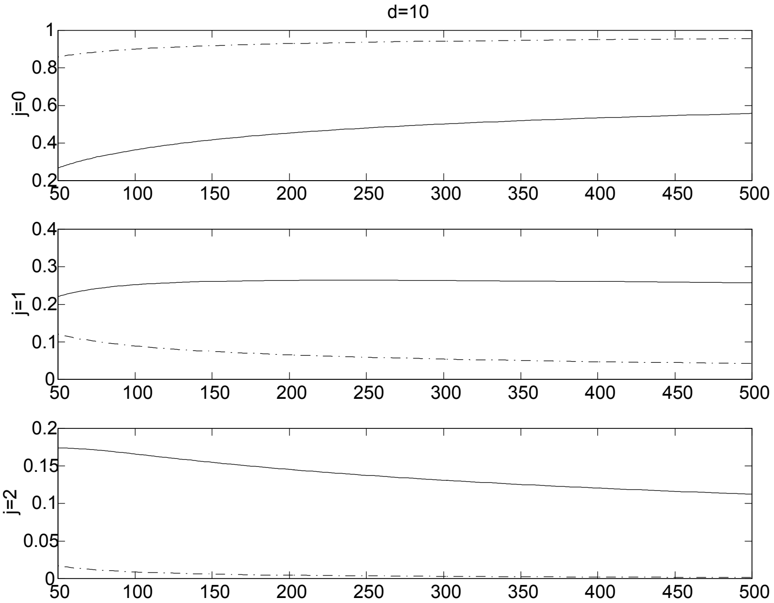

Figure 2: Standardized penalty of BIC method - dash-dot line-, and FBF method -solid line- where the maximum degree d is 10, for the models constant, $j = 0$, linear, $j = 1$, and quadratic $j = 2$.

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

Table 1: Model used in the Monte Carlo study. In all the six cases the distribution of the error term is N(0,Iσ^2)

4 Simulations

In this section we compare, using a Monte Carlo study, the procedures presented in the previous section. In addition, we study the effect of the prior distribution on these procedures using the two priors defined by (7) and (8). We envisage the following scenario: we generate observations by using a model M_j of Table (1), (see Figure 3 for a sample generated from each one of these models), where the x values are equally spaced in the interval $[-3, 3]$ so that the sample size in each case is $n = 61$. $N = 100$ replications have been generated from each model. For each replication, polynomial models of order $0, 1, ..., d = p + h$, where p is the correct degree of each model and $h = 0, 1, 2, 3, 4$, are fitted and the posterior probabilities for each possible order are computed by using the previous eight methods described in section 3.5.

The number of training samples used in the CD method is $NT = \min(100, T_0)$ and T_0 satisfies (14) for $\delta = 10^{-3}$. The value of NT depends of the number d of different models whose posterior probability we want to estimate and broadly increases linearly with the number of regessors. The empirical relationship found in this case between NT and d is shown in Figure 4.

4.1 Posterior probabilities of the models

Table 2 shows the posterior probability for the correct degree for each of the eight methods . We emphasized, highlighting in bold type, the maximal posterior probability for each model and for each value of d. The first three models present small uncertainty about the correct degree and for them the posterior probabilities of the correct degree are very high for all the methods. The last three models show more uncertainty and only for $h = 0$ the posterior probability with some of the methods is higher than 0.9.

When h increases, the posterior probabilities for the correct degree decreases, as expected. With regard to the prior, for models of small degree (see $M_1 - M_3$) the probabilities are higher with the informative prior whereas we this effect changes for higher order models (for instance, compare model 1 and model 6). With respect to the four methods compared, there are some differences between model of low order, as the first two models

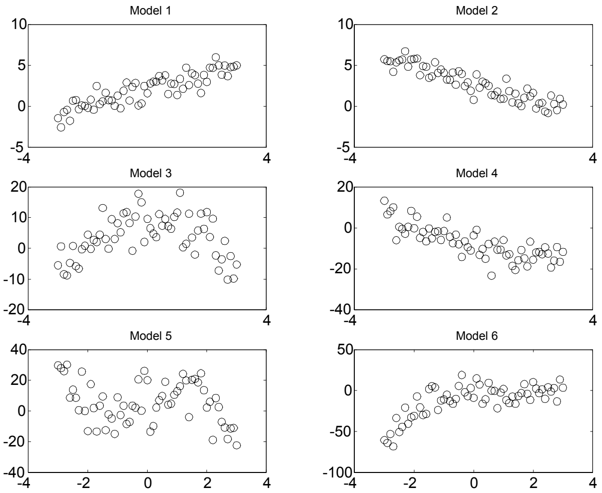

Figure 3: One replication for the six models under study.

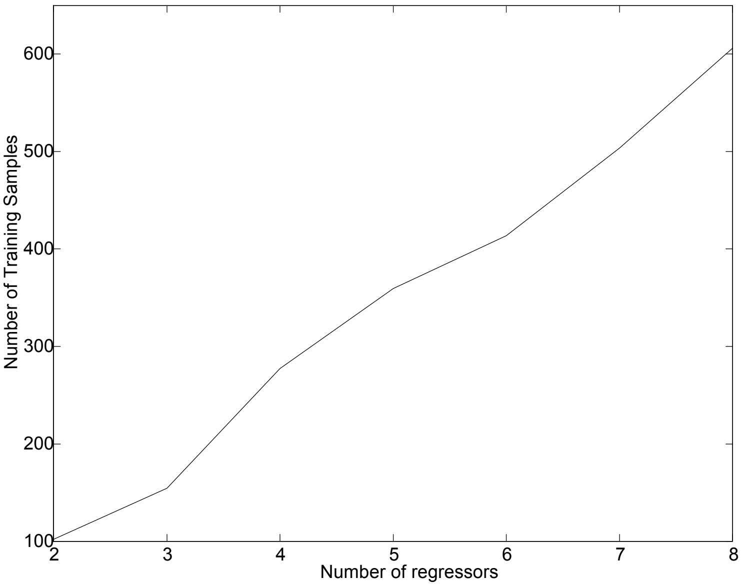

Figure 4: Number of Training Samples used to estimate the posterior probability and number of regressors in the models.

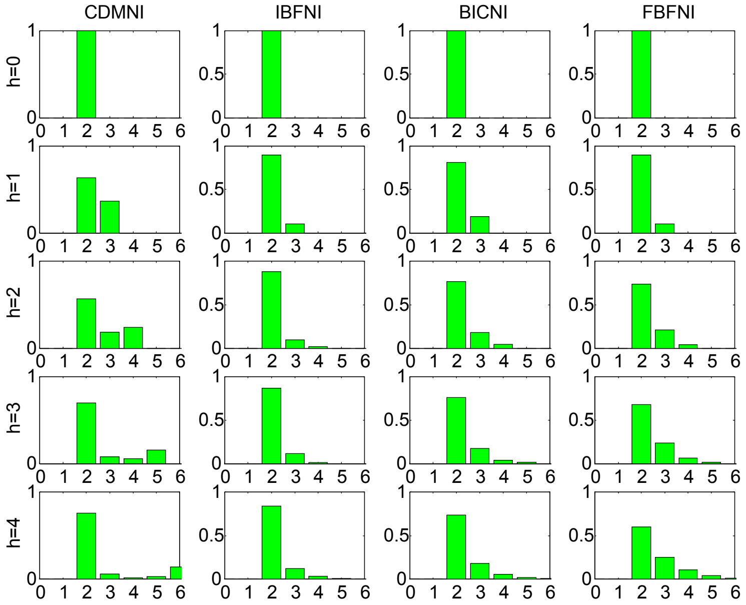

Figure 5: Posterior probabilities for different degrees in the model 3, computed with non informative prior, when h grows from 0 to 4.

that are linear and the last two models that are of third degree. For the low order models IBFI is the best method, and FBFI and BICI also seem to work very well. For the last two BICI and FBFI show similar results. From the point of view of robustness to the value of h the two methods less affected are the BIC and the IBF and this important property and their good overall performance lead us to recommend them.

In order to understand better the properties of the four methods Figure (5) shows the distribution of the values of the posterior probabilities computed for model 3, for all the possible degrees, $j = 0, \ldots, 2+h$, for five different values of h ($h = 0, 1, 2, 3, 4$). Because of space limitations, Figure (5) presents only the results for model 3 when the non informative prior is used, but the patterns are similar for other models. First note that IBF and BIC are the most robust to the value of h, a fact suggested in Table 2. Second, note that IBF penalizes more than BIC the degree of the polynomial, and BIC more than FBF, as it is to be expected from the analysis in section 3.5. The CDM shows a less stable pattern and the highest probabilities correspond first to the correct model and second to the model of maximum order included in the analysis.

4.2 Predictive Distributions

In order to compare the prediction ability of the four methods, we have generated observations at ten points equally spaced in the interval $x_h = [-3, 3]$, and have computed the response at these points from the six models described in Table 1. The process is repeated

M_1	h	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.9572	0.7154	**0.9916**	0.9224	0.9725	0.7971	0.9788	0.8283
	2	0.9698	0.6923	**0.9939**	0.9291	0.9770	0.7630	0.9734	0.7363
	3	0.9894	0.7884	**0.9939**	0.9118	0.9790	0.7356	0.9691	0.6511
	4	0.9911	0.8559	**0.9921**	0.9067	0.9723	0.7393	0.9568	0.6048
M_2	h	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
	0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	1	0.9578	0.7130	**0.9936**	0.9321	0.9736	0.7911	0.9799	0.8232
	2	0.9771	0.7276	**0.9951**	0.9375	0.9835	0.8008	0.9799	0.7720
	3	0.9902	0.8069	**0.9945**	0.9269	0.9798	0.7608	0.9702	0.6739
	4	0.9896	0.8621	**0.9901**	0.9101	0.9701	0.7518	0.9557	0.6155
M_3	h	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
	0	0.9960	0.9999	0.9937	0.9999	0.9983	**1.0000**	0.9971	**1.0000**
	1	0.9292	0.6349	**0.9835**	0.8906	0.9710	0.8082	0.9780	0.8364
	2	0.9291	0.5691	**0.9885**	0.8807	0.9744	0.7662	0.9712	0.7399
	3	0.9563	0.6980	**0.9876**	0.8688	0.9755	0.7633	0.9660	0.6795
	4	0.9526	0.7573	**0.9843**	0.8399	0.9707	0.7351	0.9559	0.5973
M_4	h	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
	0	0.8635	0.9654	0.7494	0.9220	0.8672	**0.9685**	0.8310	0.9580
	1	0.7999	0.6034	0.8040	**0.8833**	0.8871	0.8006	0.8778	0.8302
	2	0.6854	0.5018	0.7790	0.8179	**0.8440**	0.7277	0.8405	0.7064
	3	0.6652	0.5806	0.7801	0.8075	**0.8301**	0.7067	0.8257	0.6345
	4	0.5264	0.6299	0.7349	0.8128	**0.8132**	0.7268	0.8090	0.5998
M_5	h	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
	0	0.7434	0.9774	0.7295	0.9741	0.8230	**0.9919**	0.7977	0.9900
	1	0.5595	0.4879	0.7570	0.8343	0.8636	0.8133	**0.8652**	0.8440
	2	0.4033	0.3671	0.7651	0.7760	0.8374	0.7375	**0.8500**	0.7151
	3	0.3452	0.4701	0.8336	0.7857	0.8887	0.7525	**0.8958**	0.6703
	4	0.2056	0.4670	0.8432	0.7736	0.8536	0.7029	**0.8582**	0.5715
M_6	h	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
	0	0.7100	**0.9214**	0.4334	0.7320	0.6530	0.8942	0.5887	0.8605
	1	0.5086	0.4218	0.4076	0.6344	0.5753	0.6833	0.5562	**0.7098**
	2	0.3995	0.3064	0.4452	**0.6675**	0.6064	0.6568	0.6008	0.6478
	3	0.3999	0.3711	0.5143	**0.7042**	0.6441	0.6769	0.6442	0.6172
	4	0.2260	0.3409	0.4127	0.6556	0.5844	**0.6664**	0.5898	0.5675

Table 2: Posterior probability of the correct degree of the model

Mod	Meth	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI	mean
M_1	BMA	**0.00**	-0.01	-0.03	-0.04	-0.01	-0.09	-0.02	-0.08	-0.03
M_1	SBM	-0.05	-0.15	-0.05	-0.11	-0.05	-0.14	-0.05	-0.30	-0.11
M_2	BMA	-0.12	-0.11	-0.12	-0.14	-0.12	-0.19	-0.16	**-0.09**	-0.13
M_2	SBM	-0.12	-0.16	-0.12	-0.16	-0.12	-0.18	-0.12	-0.31	-0.16
M_3	BMA	**-0.59**	**-0.59**	-0.65	-0.66	-0.67	-0.67	-0.65	-0.66	-0.64
M_3	SBM	-0.64	-0.84	-0.66	-0.76	-0.64	-0.76	-0.64	-0.90	-0.73
M_4	BMA	**-0.30**	-0.38	-0.56	-0.47	-0.60	-0.63	-0.53	-0.64	-0.51
M_4	SBM	-1.12	-0.92	-0.97	-0.91	-0.90	-0.78	-0.82	-0.92	-0.92
M_5	BMA	-1.26	-1.04	**-1.02**	-1.36	-1.57	-1.41	-1.06	-1.50	-1.28
M_5	SBM	-4.55	-2.48	-2.60	-1.56	-3.31	-1.61	-2.09	-1.60	-2.48
M_6	BMA	**-0.06**	-0.10	-0.32	-0.15	-0.53	-0.37	-0.41	-0.31	-0.28
M_6	SBM	-1.23	-0.95	-1.16	-0.88	-1.25	-0.87	-1.12	-0.59	-1.01
	mean BMA	-0.39	-0.37	-0.45	-0.47	-0.58	-0.56	-0.47	-0.54	-0.48
	mean SBM	-1.29	-0.92	-0.93	-0.73	-1.05	-0.72	-0.81	-0.77	-0.90
	mean	-0.84	-0.64	-0.69	-0.60	-0.81	-0.64	-0.64	-0.66	-0.69

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

100 times, and the frequency in which the true values are included in the $85\%, 90\%, 95\%$, 97.5% and 99% highest predictive density interval (HPDI) obtained by the eight methods considered is recorded. The HPDI have been selected for each method by (i) Bayesian Model Averaging (BMA) and (ii) by Selecting the Best Model (SBM).

Let $f(\alpha, i, j)$ be the relative frequency in which the true value is included in the $HPDI(\alpha, i, j)$ interval, with $\alpha = (0.85, 0.90, 0.95, 0.975, 0.99)$, $i = CDM, IBF, BIC, FBF$ and $j = I; NI$. Let

$$d(\alpha, i, j) = (f(\alpha, i, j) - \alpha) 100$$

be the percentage deviation between the observed interval coverage and the theoretical one. Table 3 presents the values of $d(\alpha, i, j)$.

We can observe that all the values are negative, which shows that all the methods underestimate the length of the true predictive interval, that is, they underestimate the uncertainty involved in forecasting. Prediction intervals generated by Bayesian model averaging have almost always better coverage than those generated by the best selected model. The difference can be quite important when the uncertainty is relatively large, as happens with M_6, in which the percent deviation with BMA is, as an average over the eight methods, half of the deviation obtained by the best model approach. Note that BMA intervals are larger than those by SMB but this property does not imply than they have better coverage. With respect to the methods, we obtain better results for the uniform prior for the model in all the methods except for FBF, which obtains very similar results for both priors for the models.

order	Model $\widehat{y} =$	$\widehat{s}_R$	$\widehat{s}_{R(4)}$
0	12.54	2.212	1.767
1	$16.05 - \underset{(-5.16)}{0.158x}$	1.420	1.012
2	$18.67 - \underset{(-3.37)}{0.437x} + \underset{(2.20)}{0.00587x^2}$	1.282	0.8410
3	$17.98 - \underset{(-0.81)}{0.312x} - \underset{(-0.015)}{0.00027x^2} + \underset{(0.35)}{0.000087x^3}$	1.319	0.8704
4	$14.43 + \underset{(-0.55)}{0.606x} - \underset{(-0.88)}{0.0073x^2} + \underset{(0.92)}{0.0023x^3} - \underset{(-0.90)}{0.000023x^4}$	1.328	0.9031

Table 4: Different polynomial models fitted to the protein content data. 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.

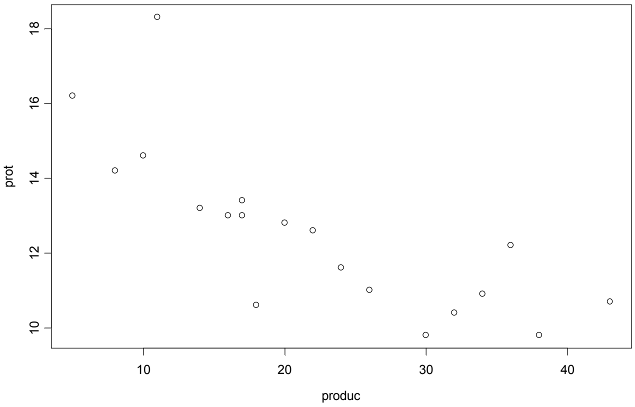

Figure 6: Graph of the Protein data

5 Examples

5.1 Protein Content

The data on wheat yield and protein content are taken from Snedecor and Cochran (1989, p. 399). This set of data has $n = 19$, and is presented graphically in Figure 5. The authors use this data to fit a quadratic model to explain the protein content given the yield. The fitted quadratic model is given in Table 4 together with other fitted models. The t value for the second order coefficient in the quadratic model is 2.20 with a p value of 0.043, that is just significant. The cubic model does not provide any improvement. As the data show, point 4 can be regarded as being “spurious”, explaining the fact that this data point is outlying. Table 4 shows the residual standard deviation for the different models fitted to both the complete data and the data set when observation 4 is deleted.

Now we compute the posterior probabilities for each order using the four procedures

d	j	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
1	0	0.0101	0.0008	0.0376	0.0029	0.0074	0.0006	0.0336	0.0026
	1	**0.9899**	**0.9992**	**0.9624**	**0.9971**	**0.9926**	**0.9994**	**0.9664**	**0.9974**
2	0	0.0856	0.0066	0.0352	0.0025	0.0061	0.0001	0.0309	0.0011
	1	0.4346	0.0745	**0.9555**	**0.8833**	**0.8195**	0.2606	**0.8696**	0.3956
	2	**0.4798**	**0.9188**	0.0092	0.1142	0.1744	**0.7393**	0.0995	**0.6034**
3	0	0.3265	0.0265	0.0335	0.0013	0.0061	0.0001	0.0371	0.0009
	1	**0.5410**	0.1047	**0.8991**	0.4877	**0.8169**	0.2203	**0.8300**	0.2790
	2	0.0841	0.1206	0.0671	**0.4945**	0.1738	**0.6249**	0.1307	**0.5856**
	3	0.0484	**0.7482**	0.0002	0.0113	0.0032	0.1547	0.0023	0.1345
4	0	**0.5332**	0.0395	0.0374	0.0020	0.0061	0.0001	0.0490	0.0010
	1	0.4430	0.1376	**0.9529**	**0.8512**	**0.8168**	0.2077	**0.8057**	0.2242
	2	0.0166	0.0287	0.0093	0.1105	0.1738	**0.5893**	0.1415	**0.5250**
	3	0.0021	0.0269	0.0001	0.0210	0.0032	0.1459	0.0038	0.1886
	4	0.0051	**0.7673**	0.0000	0.0106	0.0001	0.0569	0.0001	0.0612

Table 5: Posterior probability of the jth order model for the Protein data.

presented in Section 3. The number of training samples is 1000 and the value for the parameter of the informative prior is $q = 0.15$. The entries of Table 5 give these posterior probabilities for the jth degree using different values of d. When the prior that penalizes the degree of the polynomial is chosen, the linear model is always selected as the best one by IBF, BIC approximations and FBF. With the non informative prior BIC and FBF always choose the quadratic models for $d \geq 3$, whereas IBF tends to choose the fisrt degree model and the behaviour of CDM is towards the highest order.

Table 6 shows the posterior probability of the jth order model when the data point 4 is deleted. For the four methods with prior (8), the highest posterior probability is for the linear model (except for the constant dimension method with $d = 2$). However, when using the uniform prior all methods except the CDM choose the quadratic model and the CDM as before choose the highest possible degree. We conclude from this example that an isolated outlier may have a significant impact in the model selection process specialy with non informative prior. The stronger structure implied by the informative prior makes the result much more robust in this case as can be seen by comparing the probabilities in Tables 5 and 6.

5.2 The Voltage data.

Montgomery and Peck (1992, p. 212) presents 41 observations on the battery voltage drop in a guided missile motor over time. The scatter plot of both variables is given in Figure 7. They fitted these data by a spline with four knots. An alternative model for these data could be a polynomial regression. Montgomery and Peck state that the cubic polynomial regression shows a pattern in the residuals. We have checked that this pattern

d	j	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
1	0	0.0022	0.0002	0.0188	0.0014	0.0014	0.0001	0.0094	0.0007
	1	**0.9978**	**0.9998**	**0.9812**	**0.9986**	**0.9986**	**0.9999**	**0.9906**	**0.9993**
2	0	0.0086	0.0001	0.0098	0.0003	0.0008	0.0000	0.0080	0.0001
	1	0.2682	0.0349	**0.9348**	**0.5620**	**0.5301**	0.0781	**0.7394**	0.1800
	2	**0.7232**	**0.9650**	0.0549	0.4377	0.4691	**0.9219**	0.2526	**0.8199**
3	0	0.1187	0.0016	0.0092	0.0002	0.0008	0.0000	0.0106	0.0001
	1	**0.5473**	0.0582	**0.8158**	0.2536	**0.5257**	0.0641	**0.6876**	0.1241
	2	0.2398	0.1887	0.1746	**0.7336**	0.4653	**0.7570**	0.2969	**0.7143**
	3	0.0942	**0.7516**	0.0003	0.0108	0.0082	0.1789	0.0050	0.1614
4	0	0.3379	0.0105	0.0084	0.0002	0.0008	0.0000	0.0159	0.0002
	1	**0.5697**	0.0764	**0.7977**	0.2347	**0.5256**	0.0615	**0.6767**	0.1054
	2	0.0762	0.0844	0.1937	**0.7577**	0.4652	**0.7263**	0.2993	**0.6217**
	3	0.0070	0.0588	0.0001	0.0071	0.0082	0.1716	0.0080	0.2207
	4	0.0092	**0.7700**	0.0000	0.0002	0.0001	0.0405	0.0001	0.0520

Table 6: Posterior probability of the jth order model for the Protein data when the point 4 is deleted

order	0	1	2	3	4	5	6
residual std	2.563	2.345	1.076	0.9335	0.2576	0.2609	0.2640

Table 7: Residual standard deviation for the Voltage data for different polynomial degrees.

disappears when fitting a polynomial of fourth degree, as shown in Figure 8. Table 7 gives the residual variance for several orders and it can be seen that the fourth order model seems to fit the data quite well.

As in the previous example, in Table 8 we present the results for values of the degree from two to six with penalization parameter $q = 0.15$. All methods choose the model of degree 4 (or the highest degree when the maximum degree d is less than four) except for the method with CDM which always chooses the highest possible degree. This example is interesting becuase it shows a good agreement of the three methods IBF, BIC and FBF in chosing a high degree polynomial even in the case in which a prior penalyzing the degree of the polynomial is selected.

6 Concluding Remarks

In this paper we have carried out a comparative study for four methods to estimate the degree of a polynomial model and to obtain HDI for prediction. The four methods are compared with two different priors. The first one penalizes the degree of the polynomial and the second one is uniform over the space of the model.

We conclude that the three methods, IBF, FBF and BIC perform better than the fourth

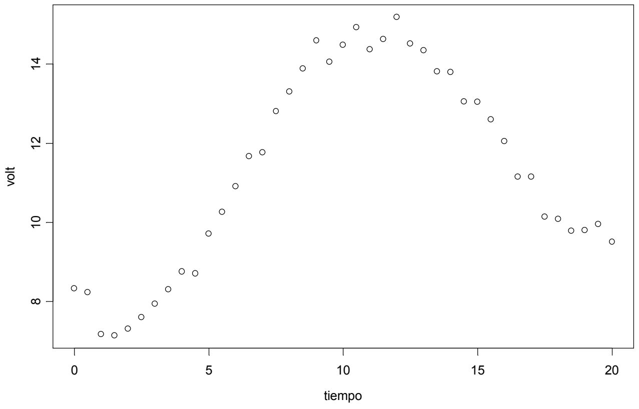

Figure 7: Graph for the Voltage data

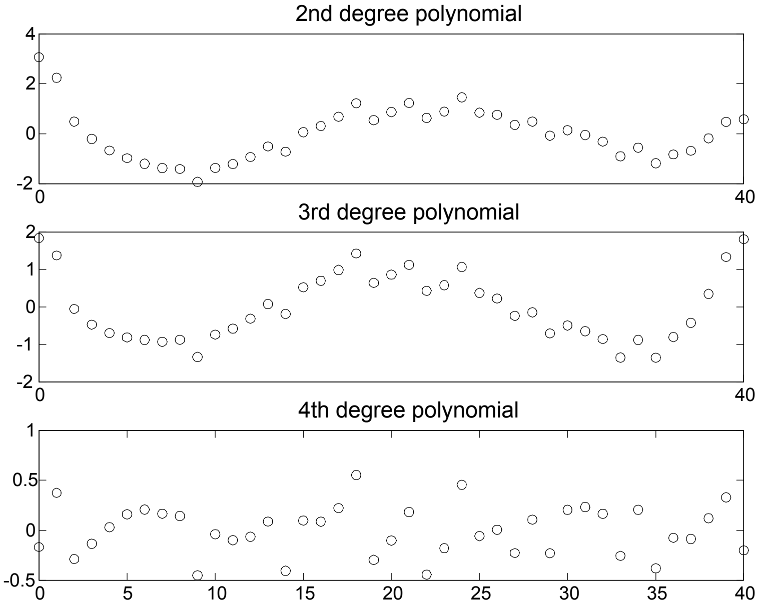

Figure 8: Residuals for different degrees of the polynomial model in the Voltage data.

d	j	CDMI	CDMNI	IBFI	IBFNI	BICI	BICNI	FBFI	FBFNI
2	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
3	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.1964	0.0476	0.0965	0.0079	0.1256	0.0107	0.2674	0.0266
	3	**0.8036**	**0.9524**	**0.9035**	**0.9921**	**0.8744**	**0.9893**	**0.7326**	**0.9734**
4	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
5	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	**0.9067**	0.4548	**1.0000**	**0.9998**	**0.9877**	**0.8571**	**0.9892**	**0.8730**
	5	0.0933	**0.5452**	0.0000	0.0002	0.0123	0.1429	0.0108	0.1270
6	0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	4	**0.8953**	0.3282	**1.0000**	**0.9998**	**0.9875**	**0.8364**	**0.9830**	**0.7918**
	5	0.0882	0.2136	0.0000	0.0002	0.0123	0.1394	0.0168	0.1807
	6	0.0165	**0.4582**	0.0000	0.0000	0.0002	0.0242	0.0002	0.0275

Table 8: Posterior probability of the jth order model for the Voltage data.

one, CD, for selecting the correct degree of the polynomial. Regarding the two priors used, broadly speaking the uniform seems to work better, although the informative prior seems to be more robust to outlier effects.

For prediction purposes, the CD and the IBF seems to provide better coverage than the more parsimonous mehods BIC and FBF. Whaterver method is used prediction intervals computed by Bayesian Model Averaging have a highest precision than those corresponding to the best model. These late intervals are underestimated, and the BMA prediction corrects somehow this effect.

References

Abramowitz, M. and Stegun, I. (1970), *Handbook of Mathematical Functions.*, Applied Mathematics Series 55, National Bureau of Standards.

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

Atkinson, A. (1978), "Posterior Probabilities for Choosing a Regression Model," *Biometrika*, 65, 39–48.

Berger, J. and Pericchi, L. (1996a), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics V*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds.). 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.

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, 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*, Plenum Publisher Corporation.

Fernández, C., Ley, E., and Steel, M. (2002), "Bayesian Modeling of Catch in a Northwest Atlantic Fishery." *Journal of the Royal Statistical Society, C. To appear.*

Geisser, S. and Eddy, W. (1979), "A Predictive Approach to Model Selection," *Journal of the Statistical Association*, 74, 153–160.

Gelfand, A. and Dey, D. (1994), "Bayesian Model Choice: Asymptotics and Exacts Calculations," *Journal of the Royal Statistical Society, Ser. B*, 56, 501–514.

Gelfand, A., Dey, D., and Chang, H. (1992), "Model Determination Using Predictive Distributions with Implementation Via Sampling-Based Methods," *Bayesian Statistics 4*, 147–167.

Gelfand, A. and Ghosh, S. (1998), "Model Choise: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11.

George, E. (1999), *Bayesian Model Selection.*, Encyclopedia of Statistical Sciences. Wiley. New York.

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*, John Wiley and Sons.

Lempers, F. (1971), *Posterior Probabilities of Alternative Linear Models*, University Press Rotterdam.

Montgomery, D. and Peck, E. (1992), *Introduction to Linear Regression Analysis*, Wiley Series in Probability, chap. Polynomial Regression Models, p. 202.

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.

San Martini, A. and Spezzaferri, F. (1984), "A Predictive Model Selection Criterion," *Journal of the Royal Statistic Society, Ser B*, 46, 296–303.

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

Snedecor, G. and Cochran, W. (1989), *Statistical Methods*, Iowa State University Press, chap. Nonlinear Relations, pp. 398–419.

Spiegelhalter, D. and Smith, F. (1982), "Bayes Factors for Linear and Log-Linear Models with Vague Prior Information," *Journal of the Royal Statistical Society, Ser B*, 44, 377–387.