AIC, BIC AND RECENT ADVANCES IN MODEL SELECTION

Arijit Chakrabarti and Jayanta K. Ghosh

OVERVIEW

As explained in e.g., [Ghosh and Samanta, 2001], model selection has somewhat different connotations in Statistics and History or Philosophy of Science. In the latter it has come to mean a major shift in paradigm on the basis of available data, of which one of the most famous examples is the shift from Newtonian Physics to Einstein's relativistic Physics on the basis of data obtained in a famous expedition of Eddington (Example 1). In Statistics it has a useful but much more pedestrian role of distinguishing between two statistical models on the basis of available data. For example, is the data coming from a normal distribution or a Cauchy distribution (see Example 5)? One of the most popular applications of Classical Statistical Model Selection is to determine which variables are important in a regression model (equivalently a linear model) for the dependent response variable y, in terms of the auxiliary variables $\boldsymbol{x}$ in the model (Example 2). However, the Classical Statistical Model Selection Rules can also be used for problems of paradigm shift (Example 1).

We use the word "Classical Statistics" to distinguish it from "Bayesian Statistics". A standard introduction to Classical Statistics is [Lehmann and Casella, 2001]. A definitive overview of Bayesian Statistics is provided in [Bernardo and Smith, 1994]. The interpretation as well as motivation of Statistical Model Selection Rules depend on whether one believes in the Classical or Bayesian paradigm of Statistics as well as the loss (or utility) function. Of the two most well-known Statistical Model Selection Rules, namely AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), AIC has a classical origin whereas BIC arises as an approximation to a Bayes rule up to $O(1)$ (the exact meaning of this statement will be explained in Section 3). At this level of approximation, one may ignore the prior distribution of the Bayesian.

By recent advances in the title we mean both recent improvements in understanding the properties and performance and relevance of AIC and BIC as well as new model selection rules and other advances. Our review of the new rules and work related to them will be brief since the AIC and BIC are our primary focus.

AIC was proposed by Akaike in a series of papers [Akaike, 1973; 1974]. He seemed to be guided by optimal prediction of a new set of y's corresponding to a

Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics.
Volume editors: Prasanta S. Bandyopadhyay and Malcolm R. Forster. General Editors: Dov M. Gabbay, Paul Thagard and John Woods.

replicate of the observed $\boldsymbol{x}$'s. It was proved first by Shibata (see [Shibata, 1981; 1983]) that in certain important problems, for large samples, AIC predicts better than any other model selection rule. He proves it by showing that asymptotically it predicts as well as an Oracle, where the Oracle is a model selection rule which always selects the best model for prediction. The prediction error of the Oracle gives a lower bound to the error committed in prediction for all model selection rules. The exact form of the Oracle considered by Shibata is given in section 2. Shibata's ideas were considerably simplified in [Li, 1987]. Interesting general results based on Li's ideas were obtained by [Shao, 1997]. We discuss more about the predictive properties of AIC in Sections 2 and 4.

BIC was introduced by Schwarz [1978] and can be used for approximating the Bayes Factor corresponding to two models and is discussed in some detail in Section 3. We briefly introduce the Bayes Factor here. Suppose M_1 and M_2 are two models specifying two families of densities $p(x|\boldsymbol{\theta_1}), \boldsymbol{\theta_1} \in \boldsymbol{\Theta}_1$ and $p(x|\boldsymbol{\theta_2}), \boldsymbol{\theta_2} \in \boldsymbol{\Theta}_2$ for the given data, where $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are the two parameter spaces corresponding to the two models. The Bayes Factor BF_{21} is the ratio of the marginal (probability) density of the data under M_2, namely $P(\text{data}|M_2)$, and that under M_1, namely, $P(\text{data}|M_1)$, with the latter in the denominator and the former in the numerator. If $BF_{21} > 1$, one chooses M_2. Otherwise one chooses M_1. A Bayesian assumes a priori probabilities for the two models under consideration and using the data posterior probabilities of the two models, namely $P(M_1|\text{data})$ and $P(M_2|\text{data})$ are obtained. BF_{21} is also equal to the ratio of the posterior odds (of model M_2 with respect to model M_1) and the prior odds. If the prior probabilities of M_1 and M_2 are taken to be half, BF_{21} is the same as posterior odds. The technical definitions of the marginal densities of data and posterior probabilities of models appear in Section 3. indexprior odds'indexHastie, T.

Both AIC and BIC are special cases of penalized likelihood rules, which may be described as follows. Suppose M_2 is a model specifying a family of densities $p(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}_2$ for the given data and M_1 is a submodel

(1) $p(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}_2.$

This is a case of nested models. (It must noted that in this case $P(M_2|\text{data}) \neq P(\boldsymbol{\theta} \in \boldsymbol{\Theta}_2|\text{data})$. We discuss this in Section 3.) The method of maximum likelihood, so popular in Classical Statistics, would suggest evaluating each model M_i by the maximized likelihood, or equivalently the maximized log-likelihood, of the data x under M_i, namely $\sup\limits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i} \log p(x|\boldsymbol{\theta})$, where "sup" denotes supremum. Since $\boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}_2$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1} \log p(x|\boldsymbol{\theta}) \leq \sup\nolimits_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_2} \log p(x|\boldsymbol{\theta}),$$

if we compare only the maximized log-likelihood we would always choose the more complex model. It is intuitively obvious that this may not be a good thing to do for all data. As pointed out in [Hastie *et al.*, 2003; Forster and Sober, 1994], a very complex model fitting the data too well is ignoring the fact that the data is composed of both signals; i.e., significant, repeatable aspects as well as noise, i.e., random perturbations of the data.

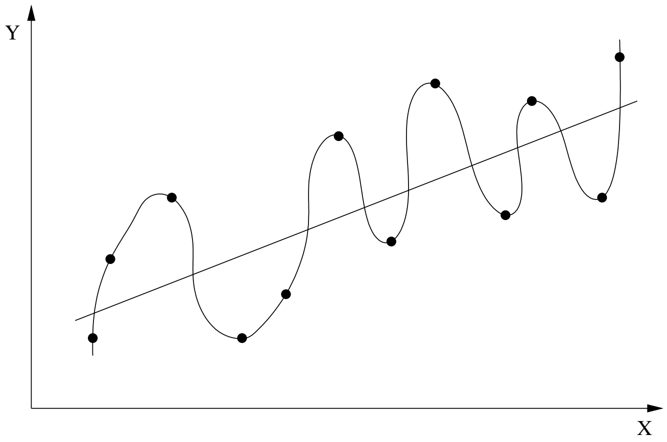

Figure 1.

As in Figure 1 above, a simpler model fitting a simple straight line is likely to be better than the zigzag curve passing through all the observed points.

This is where the so-called principle of parsimony or Ockham's Razor comes in. It suggests one should penalize each model according to its complexity. Consider the simple situation when $\Theta_2 = \mathbf{R}^{d_1}$ and $\Theta_1 = \mathbf{R}^{d_2}$ for some positive integers d_2 and d_1 such that $d_1 < d_2$. In such a case, the usual dimension of the parameter points belonging to each individual model, i.e, d_1 or d_2 is a simple measure of complexity of the model. This is so since larger the dimension, richer is the model in the sense of having a larger number of independently varying parameter components, and more complex it is to make inference on them.

This discussion leads to the penalized log-likelihood

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_i} \log p(x|\boldsymbol{\theta}) - \lambda d_i,$$

where d_i is the dimension of $\boldsymbol{\Theta_i}$ and λ is a positive constant specifying the penalty per unit dimension. If $\lambda = 1$, we get AIC. If $\lambda = \frac{\log n}{2}$, we get BIC. We choose the model which maximizes the penalized log-likelihood. For the rest of the paper, we will use the notations p or p_i to denote model dimensions.

For a long time it has been unclear which of AIC and BIC or some other penalized log-likelihood criterion is appropriate. Till very recently, there has not been much effort to understand the penalized likelihood methods in a unified manner, which would provide formal theoretical arguments in favor of each type of penalty for certain specific purposes and tell how and where each penalty comes from a particular mathematical principle. The situation is somewhat clearer now (see Section 4) with respect to AIC and BIC. Rapid progress in understanding is taking place for other rules also (see Section 5). This understanding in the case of AIC is similar to that of [Forster and Sober, 1994] and we also don't believe that AIC is a solution to all model selection problems. It is worth mentioning in this context that in a conference held in Oberwolfach in 2005 (which is also mentioned in Section 5), serious research aimed at a unified understanding of penalized likelihood methods were presented.

In Section 4 we suggest AIC and BIC are appropriate in somewhat different problems, the difference is due to different purposes and different loss functions.

Section 5 contains a brief review of some recent advances.

1 EXAMPLES

In this section we present several examples where model selection techniques can be applied to answer scientific or statistical questions. Examples 1 through 4 appeared in [Ghosh and Samanta, 2001], with appropriate references to sources of the data.

EXAMPLE 1 Eddington's experiment. According to Einstein's theory of gravitation, light gets deflected by gravitation and the amount of such deflection can also be specified. More specifically, Einstein famously predicted that under the gravitational attraction of the Sun, the light emanating from nearby stars will get deflected, but such an effect would only be visible during a total solar eclipse (when such deflection can be measured through apparent change in a star's position). To verify this prediction, a famous experiment was conducted by a team led by British astrophysicist Eddington immediately after the first world war. Four observations were collected on the amount of angular deflection (measured in seconds) by Eddington's team and other groups (spread over a period of 10 years) and those turned out to be $X_1 = 1.98$, $X_2 = 1.61$, $X_3 = 1.18$ and $X_4 = 2.24$. Einstein predicted that the amount of deflection would be 1.75. Suppose we assume that the X_i's are independently normally distributed about the unknown mean μ, i.e $X_i \sim N(\mu, \sigma^2)$, where σ^2 is unknown. To statistically test if Einstein's conjecture were true based on the observed data, we can consider choosing between the two models $M_1 : \mu = 1.75$ and $M_2 : \mu \neq 1.75$ using some model selection techniques. It is also possible to formulate this as a problem of choosing one of two nested models, in which case M_2 would permit μ to have all possible real values. Since that leads to some subtle logical questions, we postpone discussion of nested models for section 3. (However examples 3 and 4 in this section are formulated as nested models). Even though σ^2 is unknown here, just to illustrate how to use these two methods, let us treat σ^2 as known and equal to the sample variance s^2. Then making the transformation $X' = \frac{X-\mu}{s}$ of the original data X, the model selection problem becomes the same as choosing between the two models $M_1 : \mu = 0$ and $M_2 : \mu$ is a non-zero real number, where μ now denotes the mean of the transformed observations. BIC is the appropriate model selection rule here since we

Table 1. Hald's Cement hardening data

x_1	x_2	x_3	x_4	y
7	26	6	60	78.6
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

want to select the true model. The values of the BIC criterion for M_1 and M_2 are -5.675 and -6.37 respectively and so BIC selects M_2. Although AIC is not really appropriate for this purpose, it is a curious fact that AIC also selects the model M_2 in this problem. So, according to both these criteria, Einstein's prediction of $\mu = 1.75$ is supported by the observed data, although the evidence is not very strong. This particular data is now only of historical importance. Much stronger confirmation of Einstein's theory has come from other experiments, see [Gardner, 1997].

EXAMPLE 2 Hald's regression data. Table 1 below presents data on heat evolved during the hardening of Portland cement and four variables (all or a subset of which) may be able to explain the amount of heat evolved. These four variables are called the explanatory variables denoted by x_1, x_2, x_3 and x_4, which measure respectively the percentage weight of four chemical compounds (which together constitute cement) and the response variable is denoted by y, which measures the total calories given-off during hardening per gram of cement after 180 days. This data set has been analyzed several times before, e.g., [Berger and Pericchi, 1995; Burnham and Anderson, 2003].

One traditional approach in statistics to deal with such data sets is to represent y in a normal linear regression model as

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i, i = 1, \ldots, n,$$

where $x_1, \ldots, x_k$ generically denote the regressors, which according to this particular model, explain the repeatable aspect of the variation in the y values and ϵ_i are independent and identically distributed as $N(0, \sigma^2)$. Here β_i's are the unknown parameters. The different models in such situations correspond to the different possible choices of regressor variables from the pool of all potential regressors. For example, for thedata in Table 1, one can consider a total of $2^4 - 1 = 15$ possible models with the following choices of regressors $\{x_1\}, \dots, \{x_4\}, \{x_1, x_2\}, \dots, \{x_3, x_4\}$, $\{x_1, x_2, x_3\}, \dots, \{x_2, x_3, x_4\}$ and $\{x_1, x_2, x_3, x_4\}$. The purpose of choosing a model or a set of models here would be to pinpoint which regressors x's seem to have the most significant causal relationship with the amount of heat evolved y. Here AIC chooses the model with regressors x_1, x_2 and x_4 while BIC chooses the model with regressors x_1 and x_2. The AIC and BIC values for each model are reported in [Ghosh and Samanta, 2001], expressed as a difference from the value for the model, which is selected by that criterion. Thus the AIC value for model $\{x_1, x_2\}$ is -0.225, while that for model $\{x_1, x_2, x_4\}$ is 0. On the other hand, the BIC value for these two models are 0 and -1.365 respectively. So AIC favors $\{x_1, x_2, x_4\}$ while BIC favors $\{x_1, x_2\}$.

EXAMPLE 3 Nested Model selection Problem and hypothesis testing. Suppose $X_1, \dots, X_n$ are independent but identically normally distributed with mean μ and variance 1. Suppose the question to a statistician is whether the mean μ is 0 or not. The statistician formulates this problem as the testing problem with the null hypothesis $H_0 : \mu = 0$ versus the alternative hypothes $H_1 : \mu \neq 0$. Note that H_0 and H_1 specifies two disjoints subsets of $\mathbf{R}$ for the unknown μ. The same problem can also be formulated as one of selecting between two models $M_0 : \mu = 0$ and $M_1 : \mu \in \mathbf{R}$. Note that here M_0 is nested within M_1. If the data supports $\mu = 0$, then it is consistent with both models but one will choose the simpler model on grounds of parsimony, which requires that of all models which explain the data equally well, one should choose the simplest model. As explained earlier in the Overview, simplicity is defined in terms of the dimension of the model. A more detailed discussion of nested models appears in Section 4.

EXAMPLE 4 ANOVA type problems and High dimensional Setup. Suppose one has p similar normal populations, each population i having a potentially different mean μ_i, $i = 1, \dots, p$. We have r observations from each population and let $n = rp$. One might be interested in knowing whether these p means are the same or not. This question be thought of as a question of choice between the two models

$$M_1 : \mu_1 = \cdots = \mu_p \text{ vs } M_2 : \mu_i\text{'s are arbitrary.}$$

This is again a nested model problem. This is called the one-way ANOVA (Analysis of Variance) problem. Stone [1979] considered the situation where r is fixed but $p \to \infty$ and hence $n = pr$ also tends to infinity and critically studied the performance of AIC and BIC. Stone showed that AIC performs better than BIC in identifying the true model in the sense that BIC chooses M_1 with probability tending to 1 under M_2 if the μ's satisfy certain conditions and p grows sufficiently fast, while AIC chooses the correct model M_2 with probability tending to 1. Data sets for which the p =dimension=number of parameters is large is called high dimensional.

High dimensional data sets appear often these days in many applications. A prime example of this is when one wants to test gene expression levels simultaneously for thousands of genes (p in this form) but has a rather small set of data points (r in this form) on each gene, each data point corresponding to an individual. The level of expression of a gene, for example could be related to its effect on some tumor or character of the individual. It has great medical significance.

EXAMPLE 5 Normal vs. Cauchy. Suppose the following observations (when placed in increasing order of magnitude) are obtained in an experiment and the question is which distribution do the data come from :
$\{-6.56759, -3.14456, -1.19043, -0.64666, -0.64624, -0.54472, -0.43171,$
$-0.34207, -0.32573, -0.31834, -0.29348, -0.19512, -0.14658, -0.12093,$
$-0.0328, 0.075277, 0.131894, 0.187061, 0.199137, 0.214316, 0.226209,$
$0.471883, 0.654485, 0.719648, 0.788128, 0.911007, 0.946036, 1.28061,$
$1.676115, 7.986715\}$.
Looking at the data, it seems very likely that the distribution from which the data set is generated is symmetric around zero. But we are not sure what the form of the distribution is, but suspect that it might either be a normal or a Cauchy with an unknown scale parameter. The normal and Cauchy distributions with the same location are often not easy to distinguish in moderate sample sizes, but in this case we have two observations in the data set which seem too far away from the other observations, which is more likely to happen in the case of a Cauchy distribution, because its tails are very thick. But in order to decide more objectively, we want to check statistically whether M_1 : The unknown distribution is a Cauchy distribution with location 0 is true or M_2 : The unknown distribution is a normal distribution with location 0 is true. We use BIC for this purpose. It chooses model M_1, the difference in BIC criterion value for the two models being 51.7069. In fact BIC chooses the correct model here since the data were simulated from a Cauchy distribution.

2 THE AKAIKE INFORMATION CRITERION (AIC)

In this section we consider the Akaike Information Criterion (AIC) in a few canonical statistical problems and state results of its statistical optimality therein. We also discuss its connection with other model selection criteria and some of the generalizations of it. The optimality is connected with Akaike's original motivation as brought out in [Forster and Sober, 1994] but it does not follow as an immediate consequence. In fact the proofs are quite non-trivial.

We start by introducing the linear model. Consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, a vector of observations of the response (dependent) variables and let $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X_p})$ be the $(n \times p)$ matrix of explanatory variables, $\boldsymbol{X_j}, j = 1, \ldots, p$ being the j-th column of $\boldsymbol{X}$. This is the same as the set up in Example 2 in Section 1, but now written in matrix notation. In the linear model, as the name suggests, one connects the mean vector $\boldsymbol{\mu} = E(\boldsymbol{Y}|\boldsymbol{X})$ (assuming that $\boldsymbol{X}$ is fixed) with the explanatory variables via the relationship $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathrm{R^p}$ is the unknown parameter of interest. It is further assumed that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_{n\times n})$ is the vector of random errors and σ^2 is its variance, which may be known or unknown. In this context a model M specifies a certain subset of the $\boldsymbol{\beta}$ vector to be equal to zero while the others are allowed to be arbitrary. We will, for simplicity, assume that the model space is $\mathcal{M} = \{M_1, \ldots, M_p\}$, where under model M_j, $\beta_k = 0$ for $k > j$ while $\beta_1, \ldots, \beta_j$ are arbitrary. This is the nested model scenario (as also seen in Examples 3 and 4), since if M_j is true then so are $M_{j+1}, \ldots, M_p$, for any any $j \in \{1, \ldots, p\}$. Model M_j postulates that only the first j explanatory variables are potentially responsible for the variability in the repeatable aspect of the $Y's$ (measured by the mean) while the others do not contribute anything. Assuming that σ^2 is known, the Akaike Information Criterion (AIC) for model M_j is

$$\mathrm{AIC}(j) = \log L(\hat{\boldsymbol{\beta}}_j) - j,$$

where $\hat{\boldsymbol{\beta}}_j$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under model M_j and $\log L(\hat{\boldsymbol{\beta}}_j)$ is the joint density of the data under model M_j evaluated at $\hat{\boldsymbol{\beta}}_j$. AIC chooses M_j which maximizes AIC(j) among $j \in \{1, 2, \ldots, p\}$. (Under the assumption that $\boldsymbol{X}'\boldsymbol{X} = I_{p\times p}$, one can easily derive this criterion using direct calculation in this setup.) Upon simplification, this criterion can be written equivalently as (ignoring constants depending on n but independent of j)

$$\mathrm{AIC}(j) = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}||^2 + 2j\sigma^2,$$

and one minimizes AIC(j) over $j \in \{1, 2, \ldots, p\}$ to choose the best model, according to this criterion. If σ^2 is unknown, AIC(j) becomes (with σ^2 estimated by its maximum likelihod estimator under model M_j),

$$\frac{n}{2}\log(2\pi) + \frac{n}{2} + \frac{n}{2}\log\left(\frac{||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}||^2}{n}\right) + (j+1).$$

An alternative method estimates σ^2 by either the maximum likelihood estimator of σ^2 under the largest model or some estimator which is consistent under all models. It can be shown that the difference between the AIC for unknown σ^2 and this form of AIC with a plug-in estimator of σ^2 is, for large sample size n, approximately a constant depending on n but independent of the j (i.e the model under consideration), under pretty mild conditions. If, instead of the nested model scenario, one considers $\mathcal{M}$ to be any collection of models (each of which specifies a certain subset of the coordinates of $\boldsymbol{\beta}$ to be zero), the definition of AIC(M) (upto constants independent of models) for a generic $M \in \mathcal{M}$ with number of free parameters p_M, becomes AIC$(M) = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}||^2 + 2p_M\sigma^2$ if σ^2 is known and AIC$(M) = \log\left(\frac{||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}||^2}{n}\right) + 2\frac{p_M}{n}$ if σ^2 is unknown.

One very important problem where AIC can be used as a model selection rule is the problem of nonparametric regression, where the functional form of dependence between the dependent variable and the regressor is not expressible in terms of finitely many unknown parameters, as, for example, in the usual polynomial regression problem where the regression function may be known to be a polynomial of degree five in the regressor, with the six coefficients of the polynomial being the unknown parameters. Instead, the nonparametric regression model states that the expected value of Y given x is some unknown $f(x)$ where f can belong to a pretty large class of functions, e.g, say, the class of all functions which are square integrable. The reason for assuming that f might belong to a very large class of functions is the general perception that the relationship could be pretty complex.

To illustrate our point, we describe a practical example where one does not really have much clue about the relationship between x and Y to start with and the use of nonparametric regression is much more appealing intuitively than the usual parametric regression. The cosmic microwave background (CMB) radiation data from the Wilkinson Microwave Anisotropy Probe (WMAP) is analyzed in several chapters of [Wasserman, 2006] using different approaches to nonparametric regression. The basic data is a temperature map obtained by the WMAP, showing the temperature in different points of the sky 13 billion years ago. The fluctuations in the temperature map, measured through the strength of temperature fluctuations $f(x)$ (called power spectrum) at each frequency x (or "multipole moment"), provide information about the early universe. So estimation of $f(.)$ is the most interesting thing to cosmologists. Through an appropriate procedure, the temperature map can be transformed to a scatterplot of *estimated* power Y versus frequency x, given by $(x_1, Y_1), \ldots, (x_n, Y_n)$. The goal of nonparametric regression will be to estimate the function f, based on the Y's with very little assumption on its functional form.

In general, based on observations $Y_i, i = 1, 2, \ldots, n$ of the dependent variable at regressor values $\{x_i, i = 1, \ldots, n\}$, one writes the nonparametric regression model as

$$Y_i = f(x_i) + \epsilon_i, i = 1, \ldots, n,$$

where f is assumed to belong to given (large) class of functions and ϵ_i are i.i.d. errors with zero mean and finite variance. If f is square integrable, one can represent the function uniquely as an expansion which is an infinite linear combination of certain sine and cosine functions (called the basis functions), appearing in a specific order. This, in mathematical parlance, is known as the Fourier expansion of the function. Each function is determined by the coefficients (called the Fourier coefficients) that multiply the basis functions in its expansion, i.e., if the Fourier coefficients of two functions are the same, then the two functions must be identical almost everywhere. So estimating f becomes same as estimating its Fourier coefficients. Since one only has finite amount of data, infinitely many Fourier coefficients can't be estimated. Noting that as one goes further down the expansion, the Fourier coefficients become increasingly negligible, one natural solution of estimating f then is to approximate it by an appropriately chosen partial sum of this expansion and the problem of choosing a partial sum becomes one of variable selection in linear regression with the basis functions as the regressors (variables to choose from), and the Fourier coefficients as the regression coefficients. This way, the problem of estimation of f also becomes one of model selection, where each model specifies which basis functions will be used to describe f. *Note that this way each model is a false model, being a finite sum of sine and cosine terms, approximating the true infinite sum.* A popular choice of models is by taking them as nested ones, with model M_k considering the partial sum involving the first k basis functions and Fourier coefficients. AIC can be defined here exactly as in the linear model setup considered earlier, and by choosing a model here one simply wants to select the correct order of the partial sum for the given sample size with the goal of good estimation of f and hence good predictive performance with the selected model. Chakrabarti and Ghosh [2006a] show that in this nested model scenario, the model selected by AIC achieves this goal by proving that the estimate (using least squares estimates of the Fourier coefficients under the model chosen by AIC) of the unknown function converges to the truth very fast, at the so-called minimax rate. We will not delve into the technical details of this statement in this paper.

Many authors have studied asymptotic optimality properties of AIC in terms of predictive performance. To sum up, this line of research shows that under some conditions, AIC is able to predict as well as an Oracle asymptotically. The Oracle provides a lower bound for predictive performance which may possibly be attained only asymptotically but can not be implemented for finite sample size since the Oracle depends on the unknown value of the parameter. For example, in the nonparametric regression problem, Shibata [1983] defined an Oracle as

$$M_n^* = \text{argmin}_{M \in \mathcal{M}_n} E||\boldsymbol{X\beta} - \boldsymbol{X\hat{\beta}_j}||^2,$$

where the $\boldsymbol{X}$ in the above expression denotes the $n \times \infty$ design matrix involving all the basis functions, β is the true Fourier coefficient of the unknown function, $\hat{\beta}_j$ is the least squares estimate of β under model M_j and $\mathcal{M}_n$ is the model space which varies with sample size. It is easy to see that M_n^* depends on the true β for each n. Shibata [1983] showed that the ratio of the risk of the model selected by AIC and M_n^* tends to 1 as $n \to \infty$, for each true β.

We are now in a position to briefly indicate Akaike's original rationale or motivation behind the definition of the criterion named after him, without going deep into the technical details. Suppose f is the true unknown density from which i.i.d sample observations $Y_1, \ldots, Y_n$ are generated. The job of the statistician is to mimic the truth based on sample data. In parametric inference one considers a set of models $\mathcal{M}$ and each model consists of densities indexed by a finite number of parameters. The goal is to find a model which is closest to the truth in the sense that it contains a density which is closest (among all densities included in all the candidate models) to the true density in terms of some appropriate measure of divergence. As a measure of divergence between the true density f and an approximating density g, Akaike considered the Kullback-Leibler divergence, given by

$$K(f,g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

(Note that for measuring closeness (or lack of it) between two densities, the Hellinger distance and the Kullback-Leibler divergence are standard measures used by the statisticians and have been used by philosophers also in a slightly different context (see [Joyce, 1999]). For each model M_k in $\mathcal{M}$, one can find the maximum likelihood estimator $\hat{\theta}_k$ based on $Y_1, \ldots, Y_n$ under that model. Then, letting $g_k(.|\hat{\theta}_k)$ as the representative density from model M_k, Akaike considered minimizing over $M_k \in \mathcal{M}$, the criterion

$$\text{(2)} \quad \frac{1}{n} E\{ \int f(\boldsymbol{Y}^{\text{new}}) \log \left(\frac{f(\boldsymbol{Y}^{\text{new}})}{g(\boldsymbol{Y}^{\text{new}}|\hat{\boldsymbol{\theta}}_k)} \right) d\boldsymbol{Y}^{\text{new}} \},$$

where the expectation is taken with respect to the true density f and $\boldsymbol{Y}^{\text{new}}$ denotes an independent sample of size n from f. This quantity in (2) measures the divergence per observation, between the predicted density and the truth. (Criterion (2) is the most general one, suitable for all situations, e.g. the linear regression setup described above. But if the Y_i's are i.i.d as in our case the criterion reduces to $E\{\int f(Y^{\text{new}}) \log \left(\frac{f(Y^{\text{new}})}{g(Y^{\text{new}}|\hat{\boldsymbol{\theta}}_k)} \right) dY^{\text{new}}\}$, where Y^{new} is a sample of size one from f.) Minimizing (2) is again equivalent to maximizing

$$\frac{1}{n} E\{ \int f(\boldsymbol{Y}^{\text{new}}) \log(g(\boldsymbol{Y}^{\text{new}}|\hat{\boldsymbol{\theta}}_k)) d\boldsymbol{Y}^{\text{new}} \},$$

with respect to k. It was shown by Akaike that in large samples, an approximately unbiased estimator of the above quantity in (2) is given by

$$\frac{1}{n}(\log(L(\hat{\boldsymbol{\theta}}_k) - \dim(M_k)),$$

where $\dim(M_k)$ is the number of free estimable parameters in model M_k.

In the linear model example, under the assumption of orthogonal design matrix, a simple calculation shows that an exactly unbiased estimator of $\frac{1}{n}E\{\int f(\boldsymbol{Y}^{\text{new}}) \log(g(\boldsymbol{Y}^{\text{new}}|\hat{\boldsymbol{\theta}}_k))d\boldsymbol{Y}^{\text{new}}\}$ is given by $\frac{1}{n}(\log(L(\hat{\boldsymbol{\theta}}_k)) - k)$.

Now we will briefly mention the connections of AIC with some other model selection criteria. Although the Akaike Information Criterion (AIC) is considered mainly as a non-Bayesian i.e., classical statistical criterion, there has been some studies which show that in certain normal linear model problems, as in Example 4, it has an Empirical Bayes interpretation (where at least some part of the prior is estimated from the data, see e.g., [Ghosh *et al.*, 2006, chapter 9]), in the sense that under some conditions, Empirical Bayes Model selection rule and AIC choose the same model either for each sample size or asymptotically. In all these problems a model is chosen so as to minimize the expected posterior loss (using least squares estimates) in prediction of a new replicate of the dependent variable at a fixed predictor value. See [Mukhopadhyay and Ghosh, 2003; Chakrabarti and Ghosh, 2007] for further details on this. It is also worth mentioning here that AIC can be also related to a relatively recent model selection criterion DIC of [Speigelhalter *et al.*, 2002]. Spiegelhalter *et al.* [2002] define, what they call, a Bayesian measure of model complexity or the effective number of parameters in a model using information theoretic considerations. DIC is then defined as a penalized version of a Bayesian measure of fit, the penalty being the model complexity. This is similar in spirit to the usual penalized likelihood model selection criteria, where a measure of fit is often measured by the (minus) twice-maximized log-likelihood. As observed in [Spiegelhalter *et al.*, 2002; Chakrabarti and Ghosh, 2006b], under the assumption of posterior normality of the parameters in the model, DIC coincides with AIC asymptotically.

Last but not the least, we would like to point out the connection of AIC with cross-validation. The cross-validatory way of model selection, as the name suggests, keeps a part of the sample for the estimation of the parameters in candidate models and uses the remaining part of the data for validation and hence choice of the appropriate model, the idea being not to use the same data twice for two different purposes. In its simplest form, namely the leave-1-out cross validation, one just keeps one observation away at a time and calculates for each candidate model, the sum of squared error prediction losses in predicting one observation based on all remaining observations by estimating the parameters in the model using them. One chooses that model which minimizes this cross-validatory criterion. Using certain regularity conditions, Stone [1977] argued that this form of cross-validation and AIC are equivalent, in the sense that the difference between the criteria values for a give model becomes negligible in large samples. But as observed in [Chakrabarti and Ghosh, 2007], this fact seemed to have been "overinterpreted" by people. It turns out that if the model under consideration is not the true model (which by the way is one of the crucial assumptions in [Stone, 1977]), then the widely believed equivalence fails to hold. Some concrete examples of this phenomenon and theoretical explanations are given in [Chakrabarti and Ghosh, 2007].

Finally, for non-nested (or subsets) model selection scenario a modification of AIC with an additional penalty seems to help choose a more parsimonious model, as studied in [Chakrabarti and Ghosh, 2007]. The modification of AIC is

$$\mathrm{AIC}(M) = \log(\hat{\theta}_M) - p_M + p_M \log(\frac{w}{1-w}),$$

where $0 < w < 1$. The use of such a modification can be partially motivated by seeing it as a penalty for the complexity of the model space, since more complex the model space, there are more comparisons to be made and there is more chance of choosing a larger dimensional model when in fact a smaller model is also true.

3 BAYES FACTOR AND BIC

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables and the models $M_1, M_2, \ldots, M_k$ specify k parametric families of densities. The model M_j specifies the density as $p(x|\boldsymbol{\theta}_j)$ where $\boldsymbol{\theta}_j$ is a parameter in $\boldsymbol{\Theta}_j$, the parameter space corresponding to model M_j. The Bayesian analyst has to provide a prior $\pi_j(\boldsymbol{\theta}_j|M_j)$ for $\boldsymbol{\theta}_j$ conditional on the assumption that M_j is true. For notational simplicity, we will henceforth drop the subscript j and denote this prior as $\pi(\boldsymbol{\theta}_j|M_j)$. Then the probability density of the data under M_j is defined, letting $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, as

$$m_j(\boldsymbol{x}) = p(\boldsymbol{x}|M_j) = \int\limits_{\boldsymbol{\Theta}_j} \prod_1^n p(x_i|\boldsymbol{\theta}_j)\pi(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j.$$

It is known that the true model belongs to this class of k models, but it is not known which one is true. The Bayesian would assign prior probability π_j for M_j to be true, where $\sum\limits_{j=1}^{k} \pi_j = 1$. We consider 0-1 loss, i.e, the loss is 0 if a true model is chosen and is 1 if a false model is chosen. The Bayes rule is to choose the model M_j, for which the posterior probability given by

$$P(M_j|\boldsymbol{x}) = \frac{m_j(\boldsymbol{x})\pi_j}{\sum m_j(\boldsymbol{x})\pi_j}$$

is the largest for the given $\boldsymbol{x}$. If the k models are assumed to be equally likely, i.e $\pi_j = \frac{1}{k}$, for $j = 1, \ldots, k$, the Bayes rule reduces to choosing M_j that maximizes $m_j(\boldsymbol{x})$ for the given $\boldsymbol{x}$.

Consider now k nested models $M_1, M_2, \ldots, M_k$ specifying $\boldsymbol{\theta} \in \boldsymbol{\Theta}_j$, where $\boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}_2 \subset \cdots \subset \boldsymbol{\Theta}_k$. We assume M_k is a true model, but we do not wish to choose it if a more parsimonious model is also true. It is assumed that under the most complex model there is a given density $p(\boldsymbol{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_k$. This leads to a similar assumption for any M_j $(j < k)$, namely, that under M_j, there is a density $p(\boldsymbol{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_j$.

We now discuss the usual assignment of probabilities to these models M_j. We distinguish logically between a model M_j and its associated $\boldsymbol{\Theta}_j$. Though the $\boldsymbol{\Theta}_j$'s are nested in the usual set theoretic sense, we do not take the models as logically equivalent to their associated parameter sets. Rather, models are hypotheses about values of θ, often they are scientific hypotheses about natural phenomena. We consider a historical example. We imagine we are contemporaries of Galileo and are speculating on

$$M_1 : \theta = 0 \qquad \text{and } M_2 : \theta \text{ is an arbitrary real number},$$

where θ is the true difference in falling times of two objects in vacuum. Notice that M_1 is Galileo's scientific hypothesis backed by his knowledge and intuition, and $\pi(\theta|M_1)$ is Galileo's conditional probability for θ, given by $\pi(\theta = 0|M_1) = 1$. On the other hand , the conditional probability that θ takes any particular value under $\pi(\theta|M_2)$, is an assignment based on no particular knowledge and so small for all θ.

The Bayesian may choose these conditional probabilities as his given that he tries to put himself in the same frame of mind as that of Galileo and others, and the background of each model. Alternatively, for each j, he may pretend that he believes in M_j and in that state of mind assign a conditional subjective probability distribution on $\mathbf{\Theta}_j$. This is admittedly difficult, if not impossible, but all we are trying to say is that $\pi(\theta|M_1)$ and $\pi(\theta|M_2)$ are logically unrelated, and that it makes sense to take, as Bayesians usually do, $\pi(\theta = 0|M_1) = 1$ and $\pi(\theta|M_2)$ is some suitable low information prior suggested by Jeffreys.

We wanted to make two points above through an illustrative example. The first is that $\pi(\boldsymbol{\theta}|M_j)$ for different M_j's are not logically determined by the assignment for the most complex model. The second is that we need the assignment $\pi(\boldsymbol{\theta}|M_j)$ to complete our Bayesian set up for selecting one of a set of nested models.

A Bayesian would also assign prior probability π_j to M_j, where $\sum\limits_{j=1}^{k} \pi_j = 1$ and, usually, because of parsimony, $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_k$. The common (so called objective) choice is $\pi_j = \frac{1}{k}$ for all $j \in \{1, \ldots, k\}$. In our historical example, this would be the Bayesian's choice if he didn't want to favour either Galileo or the general public. We emphasize again that the Bayesian does not logically identify the model with the associated parameter space and thus can still assign equal probability to two nested models indicating his degree of belief about the truth of two competing models. The conditional density of $\boldsymbol{\theta}$ given M_j usually assigns much more probability to $\mathbf{\Theta}_j \cap \mathbf{\Theta}_{j-1}^c$ than to $\mathbf{\Theta}_{j-1}$. In fact usually, the conditional probability of $\mathbf{\Theta}_{j-1}$ is zero in most actual problems, because $\mathbf{\Theta}_{j-1}$ has usually lower dimension than $\mathbf{\Theta}_j$.

Combining all the components we can calculate the conditional or posterior probability of M_j given x as

$$p(M_j|\boldsymbol{x}) = \frac{\pi_j \int\limits_{\mathbf{\Theta}_j} p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|M_j)\, d\boldsymbol{\theta}}{\sum\limits_{1}^{k} \pi'_j \int\limits_{\mathbf{\Theta}_j{}'} p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|M'_j)\, d\boldsymbol{\theta}}.$$

The Bayes rule is to choose the model with maximum posterior probability. In the usual case where $\pi_j = \frac{1}{k}$, $j = 1, \ldots, k$, this is the same as choosing a model that maximizes

$$\int\limits_{\mathbf{\Theta}_j} p(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|M_j)\, d\boldsymbol{\theta}$$

which is usually called the marginal density of x obtained by integrating out $\boldsymbol{\theta}$. Note that since the $\pi(\boldsymbol{\theta}|M_j)$'s are different in the sense that one is not logically derived from the other, there is no reason to expect that the method will always choose the most complex model. Under regularity conditions (see e.g. [Schwarz, 1978; Ghosh *et al.*, 2006, Chapter 4]), logarithm of the above marginal likelihood is

$$\mathrm{BIC} + O(1)$$

where BIC (Bayes Information Criterion) $= \log L(\hat{\boldsymbol{\theta}}_j) - \frac{p_j}{2}\log n$, where p_j is the dimension of $\boldsymbol{\Theta}_j$, $\hat{\boldsymbol{\theta}}_j$ is the maximum likelihood estimator of θ under model M_j and $L(\hat{\boldsymbol{\theta}}_j)$ is the joint density of $x_1, \ldots, x_n$ under model M_j for $\theta = \hat{\boldsymbol{\theta}}_j$.

One often assigns probability in a different way in the non-nested case. Let $\pi(\boldsymbol{\theta})$ be a probability density over $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \cup \ldots \cup \boldsymbol{\Theta}_k$ and $p(\boldsymbol{x}|\boldsymbol{\theta})$ the density of $\boldsymbol{x}$ under $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In this case M_j's do correspond to just the subset $\boldsymbol{\Theta}_j$ specified by M_j and the p_j's are determined automatically from this identification as $\pi_j = \int_{\boldsymbol{\Theta}_j} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. It obtains from a simple algebra that the posterior probability $P(M_j|\boldsymbol{x}) = P(\boldsymbol{\theta} \in M_j|\boldsymbol{x})$ is given by

$$P(M_j|\boldsymbol{x}) = \frac{\int_{\boldsymbol{\Theta}_j} p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

In this assignment of probabilities, $P(M_j'|x) \le P(M_j|x)$ if for some pair $\boldsymbol{\Theta}_{j'} \subset \boldsymbol{\Theta}_j$. For nested models with $\boldsymbol{\Theta}_j$'s of different dimension, any assignment of probabilities of this kind to lower dimensional sets would lead to zero prior and posterior probability. More importantly, this assignment ignores the possibility of the density of $\boldsymbol{\theta}$ depending on the model. In the nested case as when θ=effect of a drug the probability distribution of θ under $M_1 : \theta = 0$ and $M_2 : \theta \in \mathbf{R}$ are often expected to be very different. We present below a concrete example to illustrate some of our points in this discussion.

EXAMPLE 6. Let θ=effect of a drug on, say, the (systolic) blood pressure. It is common to test $M_1 : \theta = 0$ (no effect) against $M_2 : \theta \le 0$ (most likely some good effect). This is question of selecting one from two nested models. A typical choice of priors is

$$\begin{aligned}
\pi(\theta = 0|M_1) &= 1 \\
\pi(\theta|M_2) &= \frac{1}{2\tau\sqrt{2\pi}} e^{-\frac{\theta^2}{2\tau^2}} \text{ if } \theta \le 0, \\
&= 0 \qquad\qquad\quad \text{otherwise}
\end{aligned}$$

and the density of a single observation x is deifined as

$$\begin{aligned}
p(x|\theta) &= N(\theta, \sigma^2), \theta = 0 \text{ under } M_1 \\
p(x|\theta) &= N(\theta, \sigma^2), \theta \le 0 \text{ under } M_2.
\end{aligned}$$

For simplicity we assume σ^2 is known and equal to one while $\tau^2 = 2$(say). Let $p_1 = p_2 = \frac{1}{2}$. Given data $\boldsymbol{x} = x_1, \ldots, x_n$ on n persons, one has

$$P(M_1|\boldsymbol{x}) = \frac{(\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n} x_i^2}}{(\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n} x_i^2} + \int\limits_{\theta\leq 0} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{2\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta}.$$

Similarly, $P(M_2|\boldsymbol{x})$ is given by

$$P(M_2|\boldsymbol{x}) = \frac{\int\limits_{\theta\leq 0} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{2\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta}{(\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n} x_i^2} + \int\limits_{\theta\leq 0} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{2\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta}.$$

If on the other hand we wish to test a new model $M_1 : \theta \leq 0$ (bad or no effect) against $M_2 : \theta \geq 0$ (good or no effect) (i.e a non-nested model choice scenario) and $\theta \sim N(0, \tau^2 = 2)$ (untruncated normal) one has

$$\begin{aligned}
P(M_1|\boldsymbol{x}) &= \frac{\int\limits_{\theta\leq 0} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta}{\int\limits_{\theta\in\mathbf{R}} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta} \quad \text{and} \\
P(M_2|\boldsymbol{x}) &= \frac{\int\limits_{\theta\geq 0} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta}{\int\limits_{\theta\in\mathbf{R}} (\frac{1}{\sqrt{2\pi}})^n e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\theta)^2} \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{\theta^2}{4}}\, d\theta}.
\end{aligned}$$

The BIC model selection rule, as defined above, chooses the M_j that maximizes BIC. It can be shown rigorously that under suitable regularity conditions, as $n \to \infty$ with k and p_j's held fixed, the BIC rule is asymptotically the same as Bayes rule. Both the Bayes rule and BIC are consistent in the sense that they choose the true model with probability tending to one.

In the nested case, the Bayes rule and BIC choose the true model with smallest dimension. Note that in the nested case if M_j is true then $M_{j+1}, \ldots, M_k$ are also true, in the sense that if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_j$, then it also belongs to $\boldsymbol{\Theta}_{j+1}$ and so on. Thus BIC is consistent with the generally accepted principle that model chosen should be as simple as possible if there are more than one model which are true or which explain data equally well. It is easy to show with examples that AIC need not do so. In fact, with $k = 2$ and 0-1 loss, model selection is equivalent to testing $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ vs $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_2$ (or $\boldsymbol{\theta} \in \boldsymbol{\Theta}_2 \cap \boldsymbol{\Theta}_1^c$). In this context AIC has a type

1 error probability α which is asymptotically greater than 0.05 in many simple examples. This makes it a very non-conservative test. This is one reason why Bayesians don't seem to like AIC. However, AIC has some good properties like predictive optimality not possessed by BIC, vide Sections 1 and 4. In short, no model selection rule will serve all purposes.

Before we end, we quote in the context of nested models, [Forster and Sober, 1994, paragraph 3 of section 7]. In this remark "LIN" refers to the class of all straight lines and "PAR" refers to the class of all parabolic curves of R^2. "The key element of any Bayesian approach is the use of Bayes' Theorem, which says that the probability of any hypothesis H given any data is proportional to its prior probability times its likelihood : $p(H/\text{Data})p(H) \times p(\text{Data}/H)$. However, it is an unalterable fact about probabilities that (PAR) is more probable than (LIN), *relative to any data you care to describe*. No matter what the likelihoods are, there is no assignment of priors consistent with probability theory that can alter the fact that $p(\text{PAR}/\text{Data}) \geq p(\text{LIN}/\text{Data})$. The reason is that (LIN) is a special case of (PAR). How, then, can Bayesians explain the fact that scientists sometimes prefer (LIN) over (PAR) ?"

Rather than trying to respond to the criticism in [Forster and Sober, 1994], we have tried to explain in this section the Bayesian practice and point of view as well as we can. Hopefully, this will help to some extent.

4 COMPARISON OF AIC AND BIC THROUGH AN EXAMPLE.

In this section, we consider AIC and BIC from a comparative point of view. Much research has been done on these two criteria. We try to summarize here (with minimum technicality) the knowledge about where these two criteria are suitabile (or otherwise), with the aid of an illustrative example.

It is quite clear now that AIC and BIC are appropriate for different purposes and have different motivations. BIC came out as an approximation (in large samples) to the Bayes rule using a 0-1 loss loss function when the dimensions of the models under study remain fixed. On the other hand, AIC was derived and proposed by Akaike as a rule, which, in large samples, does well in predicting an independent future set of observations based on data at hand, the prediction accuracy being achieved by minimizing the expected Kullback-Leibler divergence between the true model and candidate models. Note that the Bayes rule for 0-1 loss corresponds to choosing the model having the largest a posteriori probability, while minimizing the expected Kullback-Leibler divergence using AIC, for normal linear models, is equivalent to minimizing the expected squared error prediction loss (in predicting an independent future set of observations). It can be said, as Ghosh [2007] points out, that the penalty of BIC is appropriate and arose from the use of the 0-1 loss while that of AIC corresponds to the squared error loss. It indeed turns out, as intuitively somewhat expected from the above discussion, that AIC is a good rule if prediction is the sole purpose and BIC is good if the scientist wants to select the correct model. AIC excels in those problems where all models in the model space are incorrect or there is at most one correct model in the model space, while BIC excels if the correct model is in the model space. If more than one model is true, AIC may not choose the simplest true model.

We will now expand on thoughts of the last paragraph with a simple example. Let us consider the model

$$y_i = \mu + \epsilon_i, i = 1, \ldots, n,$$

where ϵ_i are identically and independently distributed $N(0,1)$ errors. We have two models $M_1 : \mu = 0$ vs $M_2 : \mu \in \mathrm{R}$. Once a model M_α ($\alpha = 1$ or 2) is selected, μ has to be estimated by the maximum likelihood estimator under that model, namely $\hat{\mu}(\alpha)$. Suppose one wants to minimize with respect to α the quantity $E_\mu(\frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i(\alpha))^2)$, where $z_i, i = 1, \ldots, n$ is a future set of n independent observations from the same distribution, and $\hat{z}_i(\alpha) = \hat{\mu}(\alpha)$ is the estimated value of z_i based on the y's and model M_α, $\alpha = 1, 2$. It can be shown that minimizing this is the same as minimizing with respect to α the quantity $E_\mu(\mu - \hat{\mu}(\alpha))^2$ or $(\mu - \hat{\mu}(\alpha))^2$ for each given $y_1, \ldots, y_n$. We will work with the quantity $A(\alpha, \mu) = E_\mu(\mu - \hat{\mu}(\alpha))^2$. Note that if the value of μ were known, one would know which one of M_1 and M_2 will minimize $A(\alpha, \mu)$ with respect to α by simply evaluationg $A(1, \mu)$ and $A(2, \mu)$. Minimizing $A(\alpha, \mu)$ with respect to α in this way yields the optimal predictive rule, called the Oracle, as

$$M_{\mathrm{ora}}(\mu) = \begin{cases} M_1 & \text{if } 0 \leq \mu^2 \leq \frac{1}{n}, \\ M_2 & \text{if } \mu^2 > \frac{1}{n}. \end{cases}$$

This way the Oracle, prescribes for each value of μ, the model which will be good for predicting that μ. But this rule depends on the unknown (true) μ and hence can't be used in practice. The expected loss for the Oracle is μ^2 if $0 \leq \mu^2 \leq \frac{1}{n}$ and $\frac{1}{n}$ if $\mu^2 > \frac{1}{n}$. The statistician wants to find a model selection rule which does almost as well as the Oracle for a moderate amount of data, and as well as the Oracle in the limit asymptotically for inifinite amount of data. Upon simple calculations, it turns out that AIC chooses M_2 if $n\bar{y}^2 > 2$ and M_1 otherwise, while BIC chooses M_2 or M_1 based on whether not $n\bar{y}^2$ exceeds $\log n$. For AIC, therefore, the loss is $(\bar{y} - \mu)^2$ if M_2 is chosen (i.e. if $n\bar{y}^2 > 2$), since $\hat{\mu}(2) = \bar{y}$, while it is μ^2 if M_1 is chosen since $\hat{\mu}(1) = 0$. The loss can be expressed simply as $(\bar{y} - \mu)^2 I_{n\bar{y}^2 > 2} + \mu^2 I_{n\bar{y}^2 > 2}$. Taking expectations, the expected loss for AIC is

$$E_\mu\{(\bar{y} - \mu)^2 I_{n\bar{y}^2 > 2}\} + \mu^2 P_\mu(n\bar{y}^2 \leq 2).$$

Similar arguments gives that for BIC the expected loss is

$$E_\mu\{(\bar{y} - \mu)^2 I_{n\bar{y}^2 > \log n}\} + \mu^2 P_\mu(n\bar{y}^2 \leq \log n).$$

Consider first the case when $\mu = 0$, i.e., M_1 is true and hence M_2 is also true. Then $M_{\mathrm{ora}}(0)$ selects M_1, AIC chooses M_1 if $n\bar{y}^2 \leq 2$ and BIC chooses M_1 if

$n\bar{y}^2 \leq \log n$. It is clear, noting that under $\mu = 0$, $n\bar{y}^2 \sim \chi^2_{(1)}$, that AIC chooses model 2 with a fixed non-zero probability (given by the probability of a chi-square random variable with 1 degree of freedom to be less than or equal to 2) for all n, while BIC chooses M_1 with probability tending to 1 as $n \to \infty$ (since the probability that a chi-square random variable with one degree of freedom being less than $\log n$ tends to 1 as $n \to \infty$). This reinforces a general fact that if two fixed dimensional nested models are true, AIC often times fails to choose the smaller model while BIC excels in choosing the parsimonious true model by penalizing the larger model heavily. For all n, the expected loss of the Oracle rule is 0 while that of AIC is $\frac{c}{n}$ where $0 < c < 1$ is a constant. The expected loss of BIC is always of a strictly larger order than $\frac{1}{n^{3/2}}$ and simultaneously of a strictly smaller order than $\frac{1}{n(\log n)^\delta}$ for any fixed positive δ, for all large enough n. While talking about orders, we are here following the convention that for two real sequences a_n and b_n, a_n is of a strictly larger order than b_n (and hence b_n of a strictly smaller order than a_n) if $a_n/b_n \to \infty$ as $n \to \infty$. Thus, in this situation the predictive performance of BIC is slightly better than that of AIC.

Now consider the case when $\mu \neq 0$, i.e only M_2 is true. Here for any large enough n the Oracle chooses M_2 and with probability tending to 1, both AIC and BIC choose M_2. It is easy to show that the ratios of the expected oracle loss and that of BIC and AIC tend to 1 as $n \to \infty$. But, in the absence of any knowledge of the true model and noting that the sample size is never infinite, a conservative approach to choose between AIC and BIC would be to see which rule minimizes, for any fixed n, the worst possible expected loss, i.e. which $\alpha \in \{\alpha_{\text{BIC}}, \alpha_{\text{AIC}}\}$ minimizes $\sup_{\mu \in \mathrm{R}} A(\alpha, \mu)$, where α_{BIC} and α_{AIC} denote, respectively, the model selected by BIC and AIC for the given sample. For the Oracle, this quantity is $\frac{1}{n}$. A simple calculation shows that for AIC this quantity is, for all large n, of the form $\frac{K_1}{n}$ for some constant $1 < K_1 < 2$ while that for BIC is of the form $\frac{K_2 \log n}{n}$ where $K_2 > 0$ is finite. Hence for BIC the worst possible rate is a factor of magnitude higher than that of AIC and the Oracle. This pathology is caused by the fact that for a given n, if μ is of the order of $\frac{\sqrt{\log n}}{\sqrt{n}}$, then BIC still chooses M_1 with non-zero probability, making the second term in the expression for the expected loss large. BIC chooses a lower dimensional model unless the increase in log-likelihood for adding a new parameter is at least $\frac{\log n}{2}$. So, for a fixed large n, for a small $\mu \sim \frac{\sqrt{\log n}}{\sqrt{n}}$, the expected loss of BIC for that μ will be a factor of magnitude larger than that of AIC.

To sum up, this example shows that the overall predictive performance of AIC is better than that of BIC in this problem, while BIC does a better job of selecting the correct model, notwithstanding its affinity to choose the smaller dimensional model because of its large penalty.

5 RECENT ADVANCES IN MODEL SELECTION

Although AIC and BIC are probably the most popular model selection criteria with specific utility (as described in detail) above, they are not the only solutions to all types of model selection problems. In recent years many other penalized likelihood model selection criteria have been proposed. For example, in cubic spline model fitting, the complexity of the model is taken care of by penalizing lack of smoothness of fitted curve. In LASSO (which penalizes the least squares criterion or the log-likelihhod criterion for normal linear models by the absolute values of the regression coefficients), one wants to select an optimum model in the presence of sparsity (i.e. when most regression coefficients are zero or close to zero). This is particularly useful in high dimensional problems. Fan and Li [2001] propose three desirable properties of a penalty function and choose a non-concave penalty which possesses optimal properties. However non-concavity makes the rule computationally inefficient.indexBIC

One can also see [Abramovich *et al.*, 2006] for another approach to handling sparse sequences, by connecting simultaneous testing of hypotheses and optimal rate of convergence. They conjectured in their paper that the modified BIC with per-parameter penalty of the form $\frac{\log(n/p)}{2}$ instead of $\frac{\log n}{2}$ would result in a good model selection rule providing optimal estimators in sparse sequences. Here p is the number of non-zero parameters in the model. It is woth pointing here that such a change in the BIC will reduce the difference between AIC and BIC criteria somewhat. This is so since in many real life problems the number of parameters p increases with sample size n (since complex data is is typically modelled with more complex models, i.e those with higher dimension) and if both n and p are large, the difference between the (per-parameter) penalties of AIC and (modified)BIC is expected to be smaller. (It is worth mentioning that the same penalty has been used earlier in [Pauler, 1998; Berger *et al.*, 2003], with the argument that the effective sample size per parameter is $\frac{n}{p}$ if the model under consideration has p free parameters.)

Much of the above theory for model selection is tied up with use of least squares estimates. If one uses Stein type shrinkage estimates, then also the notion of complexity changes a lot. This aspect has not been studied in the literature.

In the field of machine learning, particularly in classification problems, people have considered penalized empirical risk minimization, and used deep results from the theory of empirical processes to derive the penalty function which is sort of an upper bound of the error of estimation, defined to be the error in estimating the best approximating model (i.e. the member in the model space closest to the true model). Vapnik [1998, Chapter 6] recommends minimizing the structural risk which is essentially a penalized risk, with the penalty arising in a somewhat different way. Vapnik [1998, Chapter 6] also provides a good introduction to model selection based on Kolmogorov complexity or some approximation to it as in [Rissanen, 1978]. Other recent contributors to these aspects in machine learning are van de Geer, Koltchinskii, Bartlett, Lugosi, etc. The interested reader may search the rich literature on this. A good starting point would be the abstracts of talks presented in a conference on model selection at Oberwolfach in 2005, which are available from the internet at the link `http://www.ems-ph.org/journals/abstract/owr/2005-02-04/2005-02-04-05.pdf`. It seems that some form unification of the theory of model selection may be possible in future integrating all these apparently disconnected facts through some more fundamental considerations.

This paper has been devoted mainly to discussing several model selection criteria, most of the emphasis being given to AIC and BIC. Typically, once a model is selected, one uses that model to do further inference on the data at hand or future data obtained from similar experiments. Instead one may use a Bayesian model average for estimation or prediction by combining the Bayes estimates under different models with weights proportional to marginal likelihoods of models. Marginal likelihood is defined in Section 3. See [Raftery *et al.*, 1997; Hoeting *et al.*, 1999] for more details. A similar idea in the classical statistical context has been proposed by Hjort and Clasekens [2003], Breiman [2001], etc. These ideas have become very popular among both Bayesians and non-Bayesians.

SUMMING UP

In the context of model selection, we discusssed AIC and BIC (briefly mentioning some of their proposed modifications) along with the original motivation behind them. We studied concrete applications and theoretical results which reinforce our current big picture summary captured in the text. The BIC is more useful in selecting a correct model while the AIC is more appropriate in finding the best model for predicting future observations. Each of these facts hold under suitable conditions mentioned in the text. We also discussed some recent advances in model selection some of which are of interest to researchers in diverse fields of the scientific spectrum.

ACKNOWLEDGEMENT

The authors would like to thank Prasanta Bandyopadhyay and an anonymous referee for valuable suggestions leading to substantial improvement of the paper. In the process, it has clarified our own understanding of the subject even though we have not been able to obtain the version of Bayesian model selection that will be wholly satisfactory to philosophers. The authors would also like to thank M. Prem Laxman Das and Sumanta Sarkar for help with latex in drawing Figure 1.

BIBLIOGRAPHY

[Abramovich *et al.*, 2006] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate, *The Annals of Statistics*

34: 584-653, 2006.

[Akaike, 1973] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle, in B. N. Petrov and C. Csaki (eds.), *Second International Symposium of Information Theory.* Budapest: Akademiai Kiado, 267-281, 1973.

[Akaike, 1974] H. Akaike. A new look at the statistical model identification, *IEEE Transactions on Automatic Control AC* 19: 716-723, 1974.

[Berger and Pericci, 1995] J. O. Berger and L. R. Pericchi. The Intrinsic Bayes Factor for Linear Models, in Bernardo, J. M. et al. (eds.), *Bayesian Statistics 5* London: Oxford University Press, 23-42, 1995.

[Berger *et al.*, 2003] J. O. Berger, J. K. Ghosh, and N. D. Mukhopadhyay. Approximations and consistency of Bayes factors as model dimension grows, *Journal of Statistical Planning and Inference* 112: 241-258, 2003.

[Bernardo and Smith, 1994] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory* Wiley: Chichester, 1994.

[Breiman, 2001] L. Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), *Statistical Science* 16: 199-231, 2001.

[Burnham and Anderson, 2003] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer: New York, 2003.

[Chakrabarti and Ghosh, 2006a] A. Chakrabarti and J. K. Ghosh. Optimality of AIC in Inference about Brownian Motion, *Annals of the Institute of Statistical Mathematics* 58: 1-20, 2006.

[Chakrabarti and Ghosh, 2006b] A. Chakrabarti and J. K. Ghosh. A generalization of BIC for the general exponential family, *Journal of Statistical Planning and Inference* 136: 2847-2872, 2006.

[Chakrabarti and Ghosh, 2007] A. Chakrabarti and J. K. Ghosh. Some aspects of Bayesian model selection for prediction (with discussion), in Bernardo, J. M. et al. (eds.), *Bayesian Statistics 8* . Oxford University Press, 51-90, 2007.

[Fan and Li, 2001] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association* 96: 1348-1360, 2001.

[Forster and Sober, 1994] M. Forster and E. Sober. , How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions, *The British Journal for the Philosophy of Science* 45: 1-35, 1994.

[Gardner, 1997] M. Gardner. *Relativity Simply Explained*, Dover: New York 1997.

[Ghosh, 2006] J. K. Ghosh. Different Role of Penalties in Penalized Likelihood Model Selection Rules - abstract of talk presented at a workshop on Multivariate Statistical Methods at the Indian Statistical Institute, 2006.

[Ghosh and Samanta, 2001] J. K. Ghosh and T. Samanta. Model selection - An overview, *Curent Science* 80: 1135-1144, 2001.

[Ghosh *et al.*, 2006] J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis: Theory and Methods* Springer: New York, 2006.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, New York: Springer, 2001.

[Hjort and Claeskens, 2003] N. L. Hjort and G. Claeskens. Frequentist Model Average Estimators, *Journal of the American Statistical Association* 98: 879-899, 2003.

[Hoeting *et al.*, 1999] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian Model Averaging, *Statistical Science* 14: 382-401, 1999.

[Joyce, 1999] J. Joyce. *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press, 1999.

[Lehmann and Casella, 2001] E. L. Lehmann and G. Casella. *Theory of Point Estimation* Springer-Verlag: New York, 2001.

[Li, 1987] K. C. Li. Asymptotic optimality for c_p, c_l, cross validation and generalized cross validation: Discrete index set, *The Annals of Statistics*, 15: 958-975, 1987.

[Mukhopadhyay, 2000] N. D. Mukhopadhyay. *Bayesian Model Selection for High Dimensional Models with Prediction Error Loss and 0-1 Loss*, Ph.D. Thesis, Purdue University, 2000.

[Mukhopadhyay and Ghosh, 2003] N. D. Mukhopadhyay and J. K. Ghosh. Parametric Empirical Bayes Model Selection - Some Theory, Methods and Simulations, in K. B. Athreya et al. (eds.), *IMS lecture notes in honor of Rabi Bhattacharya*, 2003.

[Pauler, 1998] D. K. Pauler. The Schwarz criterion and related methods for normal linear models, *Biometrika* 85: 13-27, 1998.

[Raftery *et al.*, 1997] A. Raftery, D. Madigan, and A. J. Hoeting. Bayesian Model Averaging for Linear Regression Models, *Journal of the American Statistical Association*, 92: 179-191, 1997.

[Rissanen, 1978] J. Rissanen. Modeling by shortest data description, *Automatica* 14: 465-471, 1978.

[Schwarz, 1978] G. Schwarz. Estimating the dimension of a model, *The Annals of Statistics* 1978: 461-464, 1978.

[Shao, 1997] J. Shao. An asymptotic theory for linear model selection, *Statistica Sinica* 7: 221-264, 1997.

[Shibata, 1981] R. Shibata. An optimal selection of regression variables, *Biometrika* 68: 45-54, 1981.

[Shibata, 1983] R. Shibata. Asymptotic mean efficiency of a selection of regression variables, *Annals of the Institute of Statistical Mathematics* 35: 415-423, 1983.

[Spiegelhalter *et al.*, 2002] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* 64: 583-649, 2002.

[Stone, 1977] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *Journal of the Royal Statistical Society, Series B, Methodological* 39: 44-47, 1977.

[Stone, 1979] M. Stone. Comments on model selection criteria of Akaike and Schwarz, *Journal of the Royal statistical Society, Series B, Methodological* 41: 276-278, 1979.

[Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*, Wiley: New York, 1998.

[Wasserman, 2006] L. Wasserman. *All of Nonparametric Statistics*, Springer, 2006.