Model selection – An overview

J. K. Ghosh and Tapas Samanta*

Indian Statistical Institute, 203 B.T. Road, Kolkata 700 035, India

We provide an introduction to some of the most well-known model selection criteria and indicate how they work in actual examples. We also provide new insights based on current research.

FOR many scientists models are synonymous with paradigms. They are models of some aspects of reality as depicted in a particular science. So the problem of choosing a model appears when that science is at the crossroads. An example of this was the situation in the twenties, when physicists had to choose between Newton's classical theory of gravitation and the theory of gravitation in Einstein's general theory of relativity. One of our examples, Example 2, illustrates this sort of problem, but most others are of a different kind. They occur all the time.

Typically, when one has to analyse data arising from complex scientific experiments or observational studies in social sciences and epidemiology, there are various aspects that are not deterministic. One way of modelling nondeterministic phenomena is through a probability model. For complex phenomena it is quite rare to have only one plausible model, instead there are several to choose from. In all such situations model selection becomes a fundamental problem.

To the extent that large data sets are increasingly common because of advances in information technology, selecting a model has tended to become an essential part of analysis of such data. They present challenging methodological, computational and theoretical problems and have led to a fast-growing literature in both statistics and computer science.

This article reviews some of the major statistical developments in this area. No previous background in model selection is assumed. The next section presents a brief background, followed by six examples, some theory and analysis of some of the examples in later sections. The last section provides some concluding remarks. The section 'State-of-the-art' is based mainly on Shao[1] and Mukhopadhyay[2].

Background

In classical statistics, also called the Neyman–Pearsonian theory, model selection is usually made at the stage of exploratory data analysis, with all subsequent statistical analysis depending on the selected model. Occasionally, but not always, there is some study of sensitivity of the subsequent analysis with respect to the selected model. However, classical statistics does not emphasize model selection. Nor does it provide for the uncertainty due to a model that is assumed by convention or selected through exploratory analysis.

Nonetheless, there are certain areas of classical statistics where model selection has played an important role, for example, linear regression and time series. In both sets of problems one asks essentially the same question – which variables in a linear relation or a linear predictor are worth keeping? This becomes a model selection problem if one identifies each set of retained variables with a model (vide Example 3).

It would be naive to expect the best results by including all the variables in one's model. One way of seeing this is to note that it violates the fundamental scientific principle of parsimony, which requires that of all the models that explain the data well, one should choose the simplest. Another justification for not choosing the most complex problem comes from a predictive approach. Typically, predictions based on models which are too complex for a given data set will do badly because even a large data set may not be large enough to provide reliable estimates of all the parameters, i.e. unknown constants appearing in the model. One would be better off with a simpler model. In other words, the complexity of a good model ought to depend on the size and complexity of the data set and there is a threshold beyond which it does not pay to add complexity. Model selectors often refer to this as a trade-off between bias and variance. Complexity helps reduce bias, but increases variability, i.e. uncertainty in estimating the parameters of a model.

A good model selection rule provides an automatic threshold for allowable complexity which has some theoretical justification and works well in practice in some easily understood sense.

Generalized linear models is another area in classical statistics where model selection is popular[3–6]. In these applications it is used as an alternative to classical testing of hypothesis, with the advantage of not having to choose

*For correspondence. (e-mail: tapas@isical.ac.in) a conventional level of significance like 5% or 1% which is often inadequate in large samples. Burnham and Anderson[7] provide applications in ecology. However, it is unclear if uncritical use of Akaike Information Criterion (AIC; see the section 'Bayes factor, BIC and AIC'), which is very popular in this area, really provides a satisfactory alternative.

A recent development favouring model selection, is the popularity of Bayesian analysis as an alternative paradigm for statistics. There are several reasons for this. A Bayesian analysis can include the uncertainty due to models in his assessment of total uncertainty. This is not easy, but doable – certainly not as difficult as in classical statistics. Moreover, testing a hypothesis, which is a common statistical problem, is no different from model selection in Bayesian analysis (Examples 1 and 2). Finally, advances in information technology and sophisticated versions of Markov Chain Monte Carlo (MCMC) (see e.g. refs 8 and 9) have made it relatively easy to implement Bayes model selection rules.

Model selection rules are also widely used by computer scientists. One major application is in finding an optimal architecture for a neural network that is trained to do a particular task. This is very similar to model selection problem for linear regression, except that the basic relations are much more complex and nonlinear[10–12]. Another major application is in data mining.

Interestingly, model selection rules used in these computer science applications are either AIC (Akaike Information Criterion) or BIC (Bayes Information Criterion), which are the two most popular statistical rules, or their close cousins. There is a wholly different theory of model selection based on Kolmogorov's theory of algorithmic complexity and developed by theoretical computer scientists, which has not been used in data mining or neural networks. A good source for this theory is Li and Vitanyi[13]. It turns out that BIC can be justified from this point of view. Rissanen[14,15] assumes a predictive framework with a logarithmic loss comparing the actual and predictive distribution. Rissanen believes this comes close to using Kolmogorov complexity. Laplace integration, as explained in the section 'Bayes factor, BIC and AIC', leads to maximizing BIC as an approximation to the Bayes rule.

One of the major problems in model selection is that for large samples the BIC and AIC usually select very different models, the AIC providing much less penalty for complexity than the BIC.

Examples

Example 1: We begin with a simple, but generic example. We have n observations $X_1, X_2, \ldots, X_n$ known to be independent, each having a normal distribution with mean μ and variance σ^2. Symbolically, $X \sim N(\mu, \sigma^2)$. We have two models

M_1: $\mu = 0$,

M_2: μ is arbitrary.

For simplicity σ^2 will be taken to be known and equal to one. Many statistical problems are of this kind, except that σ^2 will be rarely known. Let us explore this a little further. Suppose $n = 1$, i.e. we have only one observation. Then it is clear that if $|X_1|$ is 'close' to zero, parsimony requires that we select M_1. Statistical theory shows for general n, we should do the same with the mean $\overline{X} = (X_1 + X_2 + \ldots, + X_n)/n$ replacing X_1. The catch is we do not know how to define when $|\overline{X}|$ is close to zero. According to BIC, you choose M_1 if

$$|\overline{X}| < \{(\log_e n)/n\}^{1/2}. \tag{1}$$

According to AIC, you choose M_1 if

$$|\overline{X}| < (2/n)\tfrac{1}{2}. \tag{2}$$

The BIC is more conservative, in that it chooses the simpler model more often.

Example 2: Einstein's theory of gravitation predicts that light is deflected by gravitation and specifies the amount of deflection. Einstein predicted that light of stars would deflect under gravitational pull of the sun on the nearby stars, but the effect would be visible only during a total solar eclipse when the deflection can be measured through apparent change in a star's position. A famous experiment by a team led by British astrophysicist Eddington, immediately after the First World War (see ref. 16), led to acceptance of Einstein's theory. Though many other better designed experiments have confirmed Einstein's theory since then, Eddington's expedition remains historically important. There are four observations, two collected in 1919 in Eddington's expedition, and two more collected by other groups in 1922 and 1929. The observations are $X_1 = 1.98$, $X_2 = 1.61$, $X_3 = 1.18$, $X_4 = 2.24$ (all in seconds as measures of angular deflection). Suppose they are normally distributed around their predicted value μ. Then $X_1, \ldots, X_4$ are independent and identically distributed as $N(\mu, \sigma^2)$. Einstein's prediction is $\mu = 1.75$. We will test the models M_1: $\mu = 1.75$ and M_2: μ is arbitrary, where σ^2 is unknown.

Example 3 (Hald's regression data): Table 1 presents a small set of data on heat evolved during the hardening of Portland cement and four variables which may be related to it[17]. This classic data set has been used by several authors; see Burnham and Anderson[7] for references. The sample size (n) is 13. The regressor variables (in per cent of the weight) are x_1 = calcium aluminate ($3CaO.Al_2O_3$), x_2 = tricalcium silicate ($3CaO.SiO_2$), x_3 = tetracalcium alumino ferrite ($4CaO.Al_2O_3.Fe_2O_3$), and x_4 = dicalcium silicate ($2CaO.SiO_2$). The response variable is y = total calories given-off during hardening per gram of cement after 180 days.

Usually such a data set is analysed using normal linear regression model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where p is the number of regressor variables in the model, $\beta_0, \beta_1, \ldots \beta_p$ are unknown parameters and ε_i's are independent errors having a $N(0, \sigma^2)$ distribution. There are a number of possible models depending on which regressor variables are kept in the model. In the section 'Calculations for the examples' we will analyse the data and choose one from this set of possible models using different model selection criteria.

Example 4: We consider all subsets regression as in Example 3, but add an orthogonality condition that simplifies things a lot. This makes the penalized likelihood criteria like AIC and BIC, as defined in the next section, as easy to interpret as in Example 1.

We consider the linear regression model (eq. (3)) and in addition assume orthogonality, i.e. $\boldsymbol{x}'_j \boldsymbol{x}_k = \Sigma_{i=1}^n x_{ji} x_{ki} = 0$ for $j \neq k$ where $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})'$, $j = 1, \ldots, p + 1$. With suitable renormalization we can assume

$$\| \boldsymbol{x}_j \| = \sqrt{\textstyle\sum_{i=1}^n x_{ji}^2} = 1.$$

A model consists of choosing the variables $\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_k}$ that really matter. The assumption of orthogonality ensures that the least squares estimate of β_j is $\boldsymbol{x}'_j \boldsymbol{y}$, irrespective of which model we are looking at. The penalized likelihood like AIC or BIC with penalty λ chooses the model that minimizes

$$\boldsymbol{y}'\boldsymbol{y} - \sum_{j=0}^{p} \hat{\beta}_j^2 + \lambda(p+1)\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is an estimate of σ^2. This is the same thing as choosing only $\boldsymbol{x}_j$'s for which the square of the least squares estimate of β_j exceeds $\lambda\hat{\sigma}^2$. In particular the $\boldsymbol{x}_j$'s for which AIC and BIC will differ are those satisfying

$$2\hat{\sigma}^2 < \hat{\beta}_j^2 < (\log n)\hat{\sigma}^2,$$

and these j's will be kept in the model by AIC, but not by BIC which is more parsimonious. This has the effect that BIC will do better in identifying the correct model when there is no signal and only noise, i.e. none of the explanatory variables matter, but on the whole, AIC will do better in prediction of future observations when many of the unknown β_j's are non zero, but too small to be pass through BIC.

Table 1. Cement hardening data with four regressor variables x_1, x_2, x_3 and x_4 and a response variable y

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

This example is a slight simplification of problems in telecommunication where the $\boldsymbol{x}_j$'s are related to wavelets of different bands and identifying them when there is no signal is important. In such cases the BIC or similar criteria would be preferred to AIC. We expect the situation to be similar even if there is no orthogonality, provided there is no high correlation between a pair of variables as in Example 3.

Example 5: This is a somewhat complicated example from ecology with a large data set and many parameters. Burnham and Anderson[7] (to be abbreviated below as B–A) generated data to mimic the experiment presented by Stromborg *et al.*[18]. The experiment is designed to study the survival effect of a pesticide administered to nestling European starlings in an island. All birds under study are leg-banded with uniquely numbered coloured bands. Half of these birds are randomly assigned to a treatment group and receive a dose of pesticide and the remaining birds are assigned to a control group. Birds in these two groups are believed to be under otherwise very similar conditions. After a 4-day period following the dosage, all birds are released. Surviving starlings are potentially resighted and resighting efforts are made on a day (say, Friday) in each of the following few weeks. The birds captured in a week are released again. For more details of the experiment and the data see B–A[7] and the references therein.

B–A generated data with 300 birds in both the treatment and control groups with 8 resighting occasions, weeks 2, . . . , 9. Thus the birds are released on 8 occasions, weeks 1, . . . , 8. Table 2 is reproduced from B–A and presents the data as a matrix ($n(i, j)$) for both the groups, where $n(i, j)$ denotes the number of birds released at week i and captured in week j for the first time after week i ($i = 1, \ldots, 8$; $j = i + 1, \ldots, 9$). It also gives the total number ($N(i)$) of birds released at week i ($i = 1, \ldots, 8$).

A product multinomial model is assumed for the given data. Each row corresponds to a multinomial distribution.

As denoted in B–A, let ϕ_{ti} and ϕ_{ci} be the conditional probabilities of survival from week i to $i + 1$ ($i = 1, \ldots, 8$) for the treatment group and the control group, respectively and p_{ti} and p_{ci} be the conditional probabilities of resighting at week i, $i = 2, \ldots, 9$. The possible models are

$$M_{r,s}\text{: } \phi_{t1} \neq \phi_{c1}, \ldots, \phi_{tr} \neq \phi_{cr}, \phi_{ti} \neq \phi_{ci}, \quad i = r + 1, \ldots, 8,$$

$$p_{t2} \neq p_{c2}, \ldots, p_{t,s+1} \neq p_{c,s+1}, p_{tj} = p_{cj}, \quad i = s + 2, \ldots, 9, \quad r = 0, \ldots, 8, s = 0, \ldots, 8.$$

For example, the hypothesis of no treatment effect corresponds to the model $M_{0,0}$. For discussion on these models and other related matters see B–A[7], where some other models employing transformations on the parameters are also considered. We, however, restrict only to the above models for the sake of simplicity.

It is to be noted that although B–A state that $n(i, j)$'s are being reported for $i = 1, \ldots, 9$, $j = 2, \ldots, 10$, they indeed report for $i = 1, \ldots, 8$, $j = 2, \ldots, 9$. Also the parameters ϕ_8 and p_9 are not separately estimable, only the product $\phi_8 p_9$ is estimable, so we treat this product as a single parameter. Therefore, e.g., for Model $M_{0,0}$, the number of estimable parameters is taken to be 15 (compared to 17 reported in B–A). However, this does not affect the AIC-differences (Δ-values) reported in B–A. An analysis of the data will be presented in section 'Calculations for the examples'.

Table 2. Starling data presented as a matrix ($n(i, j)$) for both the treatment and control groups where ($n(i, j)$) denotes the number of birds first captured in week j after last being released at time i ($i = 1, \ldots, 8$, $j = i + 1, \ldots, 9$) and the total number ($N(i)$) of birds released at week i ($i = 1, \ldots, 8$)

		$n(i, j)$							
Week	$N(i)$	$j = 2$	3	4	5	6	7	8	9
Recapture for treatment group									
1	300	158	43	15	5	0	0	0	0
2	158		82	23	7	1	1	0	0
3	125			69	17	6	1	0	0
4	107				76	8	2	0	0
5	105					67	20	3	0
6	82						57	14	1
7	81							53	12
8	70								46
Recapture for control group									
1	300	210	38	5	1	0	0	0	0
2	210		157	20	8	2	0	0	0
3	195			138	24	2	1	0	0
4	163				112	24	2	0	0
5	145					111	16	6	0
6	139						105	16	4
7	124							93	12
8	115								89

Example 6 (a slightly modified version of a problem of Stone[19]): We have p normal populations $N(\mu_1, 1), \ldots, N(\mu_p, 1)$. From each of them, r samples have been drawn, yielding $n = pr$ independent random variables with $X_{i1}, \ldots, X_{ir}$ distributed as $N(\mu_i, 1)$, $i = 1, \ldots, p$. One has to choose one of two models:

M_1: $\mu_1 = \ldots = \mu_p = 0$,

M_2: μ_i's are arbitrary.

We consider a situation where r is fixed, but $p \to \infty$, so that $n = pr$ also tends to infinity. Following Stone one can show AIC performs better in identifying the true model than BIC. However, Mukhopadhyay[2] has shown that in this problem BIC is not an accurate approximation to the Bayes factor, which is not surprising since Schwarz's basic assumption of fixed p does not hold there. The section 'State-of-the-art' continues a discussion of this example.

Bayes factor, BIC and AIC

We begin with the generic Example 1 to define, motivate and explore Bayes factor (BF) and BIC.

Since M_2 involves an unknown parameter μ, a Bayesian introduces a prior probability density $p(\mu \mid M_2) \equiv p(\mu)$ to integrate out μ and get the integrated likelihood of the data as

$$\int_{IR} \prod_{i=1}^{n} f(X_i \mid \mu, M_2)\, p(\mu)\, d\mu \overset{\text{def}}{=} f(X_1, \ldots, X_n \mid M_2), \tag{4}$$

where

$$f(X_i \mid \mu, M_2) = \frac{1}{\sqrt{2\pi}} e^{-(X_i - \mu)^2/2},$$

is the density of X_i under μ. The likelihood under M_1 is

$$\prod_{i=1}^{n} \{1/\sqrt{2\pi}\}\, e^{-X_i^2/2} \overset{\text{def}}{=} f(X_1, \ldots, X_n \mid M_1). \tag{5}$$

The Bayes factor $\mathrm{BF}_{1,2}$ of M_1 relative to M_2 is the ratio

$$\frac{f(X_1 \cdots, X_n \mid M_1)}{f(X_1 \cdots, X_n \mid M_2)}. \tag{6}$$

Large values of $\mathrm{BF}_{1,2}$ indicate evidence in favour of M_1, whereas small values suggest M_2 is likely to be true.

If one is also able to assign prior probabilities π_1 and $\pi_2 = 1 - \pi_1$ to M_1 and M_2, then the posterior or conditional probability of, say, M_1 given the data $X_1, \ldots, X_n$ is

$$\frac{\pi_1 f(X_1,\cdots,X_n \mid M_1)}{\pi_1 f(X_1,\cdots,X_n \mid M_1)+\pi_2 f(X_1,\cdots,X_n \mid M_2)}$$

$$=\frac{\pi_1 \mathrm{BF}_{1,2}}{\pi_1 \mathrm{BF}_{1,2}+\pi_2}.$$

One would choose M_1 if this posterior probability is greater than half, i.e.

$$\mathrm{BF}_{1,2} > \frac{\pi_2}{\pi_1}.$$

Often one sets $\pi_1 = \pi_2 = \frac{1}{2}$, in which case the criterion for choosing M_1 becomes

$$\mathrm{BF}_{1,2} > 1.$$

There is a catch in this simple machinery. One has to specify the prior $p(\mu \mid M_2)$ in order to calculate $\mathrm{BF}_{1,2}$. A common nonsubjective choice is $p(\mu \mid M_2) = 1$ or somewhat more generally

$$p(\mu \mid M_2) = c > 0. \qquad (7)$$

While this specification works well in estimation problems, there are serious problems in testing or model selection. These were first pointed out by Jeffreys[20] and recently, have been the starting point of a new theory that will be briefly discussed in the next section. There is a way of bypassing this choice by using an approximation due to Schwarz[21].

Let the integrand

$$\prod_{i=1}^{n} f(X_i \mid \mu, M_2) \equiv f(\underset{\sim}{X} \mid \mu, M_2)$$

be maximized at $\mu = \hat{\mu}$. Easy calculation shows $\hat{\mu} = \bar{X}$, where $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$ and

$$\log f(\underset{\sim}{X} \mid \mu, M_2) = \log f(\underset{\sim}{X} \mid \hat{\mu}, M_2) - \frac{n}{2}(\bar{X} - \mu)^2.$$

A famous principle of approximation due to Laplace suggests that the integral in eq. (4) is well approximated by the integral in a small range $\hat{\mu} - \delta$ to $\hat{\mu} + \delta$. Assume $p(\mu)$ is continuous and positive everywhere. On such a set $p(\mu)$ is nearly a constant, so that eq. (4) can be approximated by

$$p(\hat{\mu}) f(\underset{\sim}{X} \mid \hat{\mu}, M_2) \int_{\bar{X}-\delta}^{\bar{X}+\delta} \exp\left(-\frac{n}{2}(\bar{X} - \mu)^2\right). \qquad (8)$$

The integral in eq. (8) converges to $\sqrt{2\pi/n}$ as $n \to \infty$. All this finally leads to the approximation

$$\log f(\underset{\sim}{X} \mid M_2) = \log f(\underset{\sim}{X} \mid \hat{\mu}, M_2) - \tfrac{1}{2}\log n$$

$$+ \log\sqrt{2\pi} + \log p(\bar{X}) + o(1).$$

It is easy to justify this rigorously using the (strong) law of large numbers.

Schwarz[21] suggested the approximation up to $-\frac{1}{2}\log n$, pointing out that this does not require a specification of the prior $p(\mu)$. The log Bayes factor is approximated by

$$\log \mathrm{BF}_{1,2} = \log f(\underset{\sim}{X} \mid M_1) - (\log f(\underset{\sim}{X} \mid \hat{\mu}, M_2) - \tfrac{1}{2}\log n)$$

$$= -\frac{n}{2}\bar{X}^2 + \tfrac{1}{2}\log n.$$

More generally, suppose we have k nested models $M_1, \ldots, M_k$ with M_i having p_i parameters $\underset{\sim}{\theta}_{p_i} = (\theta_1, \ldots, \theta_{p_i})$. Then the integrated likelihood under M_i is approximated in the logarithmic scale by

$$\log f(\underset{\sim}{X} \mid M_i) = \log f(\underset{\sim}{X} \mid \hat{\boldsymbol{\theta}}_{p_i}, M_i) - \frac{p_i}{2}\log n. \qquad (9)$$

An elegant and rigorous justification was given by Schwarz[21] for what are called linear exponential families. A derivation under general regularity conditions and with p_i's fixed as $n \to \infty$ is given in Chapter 1 of Ghosh and Ramamoorthi (manuscript under preparation). The expression above is called the BIC. One chooses the model for which eq. (9) is a maximum. This is an approximation to the rule which chooses a model maximizing the integrated likelihood. Usually one multiplies eq. (9) by – 2 to get

$$-2\log f(\underset{\sim}{X} \mid \hat{\theta}_{p_i}, M_i) + p_i \log n, \qquad (10)$$

probably because the first term has a χ^2-distribution with p_i degrees of freedom, if M_i is the true model and selects the model for which eq. (10) is minimum. The χ^2-distribution will not play any role in the sequel, but *we will follow the convention that BIC is defined by eq. (10).*

The term $p_i \log n$ is a penalty for dimension p_i which measures the complexity of M_i. The term $\log n$ is called the penalty.

What makes the BIC very attractive is that it comes up with an automatic penalty that is easy to justify in some sense. AIC appeared earlier[22,23] with another automatic, but very different penalty,

$$\mathrm{AIC} = -2\log f(\underset{\sim}{X} \mid \hat{\theta}_{p_i}, M_i) + 2p_i. \qquad (11)$$

According to Akaike one chooses the model which minimizes AIC. Since the penalty is now only 2, instead of $\log n$, AIC will tend to choose much more complex models than BIC.

Which rule should one use in a particular case? Unfortunately, to answer this one needs to understand the motivation and justification for using AIC, but that is not as easy as in the case of BIC. We offer below some basic insights due to Akaike. A more precise theoretical justification will be given in the next section.

Akaike pointed out that as $n \to \infty$, in many cases the complexity of models will grow with n, i.e. the p_i's will also tend to infinity as $n \to \infty$. Moreover, the object may be to make a good prediction rather than decide which model is true. Indeed, the true model may be too complex to be included in the space of usable models. As George Box has observed, all models are false, some are useful.

Akaike suggested that in situations like these his methods should do better than BIC. On the other hand, one expects that if one has a set of fixed models, at least one of which must be true and can have as large a data set as one wants, i.e. p_i's are fixed and $n \to \infty$, then one would expect the BIC to lead to the most parsimonious true model but AIC may fail to do that. In problems like Example 1, BIC is expected to lead to the true model, but AIC may choose M_2 even if $\mu = 0$. Notice that if $\mu = 0$, logically both M_1 and M_2 are true, but parsimony requires we choose M_1 in preference to M_2. A precise justification of some of these results (due to Shao[1]) appear in the next section.

Stone[24] (discussion of Shao[1]) suggests BIC is suitable for 'hard science' with given fixed models whereas AIC is suitable for 'soft science' with models chosen for good prediction rather than discovery of the truth.

It turns out that the problem has a rather simple resolution at least in the case of our second generic problem, Example 6, vide Mukhopadhyay[2].

The next section contains a brief introduction to the results of Shao[1] and Mukhopadhyay[2].

State-of-the-art

There are three subsections. The first presents recent results on optimality of AIC in a predictive framework. The second subsection discusses some problems with nonsubjective priors like that presented in eq. (7) and various solutions proposed by Jeffreys[20] and more recently by Berger and Pericchi[25,26]. The third subsection discusses the relation between Bayes rules, including the Bayes factor and their relation to AIC as well as possibilities of improvement over AIC. The first subsection is based on Jun Shao[1], the second on Berger and Pericchi[25,26] and Ghosh and Samanta[27], the third on Mukhopadhyay[2].

Predictive optimality of AIC

Consider linear models with normal error. All the examples, except Example 5, can be put in this form. Possibly, even Example 5 can be put in this form with suitable approximations. Let

$$Y_j = \mu_j + \varepsilon_j,\ j = 1, \ldots, n,$$

where ε_j's are independent $N(0, \sigma^2)$ variables and μ_j's are unknown constants, i.e. what we have been calling parameters. Asymptotically, it does not matter much whether σ^2 is known or not – wherever σ^2 appears below, one simply substitutes an estimate that converges to σ^2 in the long run. So for simplicity we assume, as in Example 1, σ^2 is known.

What are the models? We assume $\underset{\sim}{\mu} = (\mu_1, \ldots, \mu_n)'$ lies in some linear space. Each linear space in IR^n corresponds to some model M_i of dimension p_i. We assume there are k models, k and p_i's can tend to infinity as $n \to \infty$. The true model, say, M may not be among the models being used. The prediction loss for assuming M_i is defined as follows. Assuming M_i, let the least squares estimate of $\underset{\sim}{\mu}$ be $\underset{\sim}{\hat{\mu}}(i)$. Imagine the Y_j's are fixed and suppose we have a new replicate $Y_1', \ldots, Y_n'$. Calculate the (conditional) expectation of the squared error prediction loss $\Sigma_j(Y_j' - \hat{\mu}_j(i))^2$ keeping the Y_j's and hence the Y_j's/ $\hat{\mu}_j(i)$'s fixed. The expectation is over all possible future replicates. A little algebra shows this prediction loss is of the form: $\Sigma_j(\hat{\mu}_j(i) - \mu_j)^2$ plus a part which does not depend on which model is being used. Here μ_j's correspond to the μ's under the true model M. This fact shows we can compare different model selection rules, using only the part that depends on model M_i, namely,

$$L(M_i, \underset{\sim}{\mu}, \underset{\sim}{Y}) = \sum_{j=1}^{n} (\hat{\mu}_j(i) - \mu_j)^2.$$

Given an actual rule like BIC or AIC or any general penalized likelihood rule, not only $\underset{\sim}{Y}$, but M_i is random. The chosen model depends on $\underset{\sim}{Y}$ in a highly nonlinear way, making risk evaluations of a rule, evaluation of expected loss over all $\underset{\sim}{Y}$, a practically impossible task, except asymptotically or through simulations. We only present the asymptotics.

In order to define optimality, we first define an oracle – something that is good to have, but requires special knowledge. If a rule does as well as such an oracle, without using its special knowledge, the rule must be optimal.

Now for definitions. Let the special knowledge consist in knowing the true values μ_j. Given this, the best model is chosen by minimizing $L(M_i, \underset{\sim}{\mu}, \underset{\sim}{Y})$ with respect to a variable model M_i. Let M_0 be the model chosen in this way with dimension p_0. Then with this knowledge, $L(M_0, \underset{\sim}{\mu}, \underset{\sim}{Y})$ is the best one can do. So if a model selection rule choosing a model $M_{\hat{i}}$ is such that, say,

$$\frac{L(M_i, \underset{\sim}{\mu}, \underset{\sim}{Y})}{L(M_0, \underset{\sim}{\mu}, \underset{\sim}{Y})} \to 1,$$

in the sense of convergence in probability or some other asymptotic sense, one would be entitled to calling $M_{\hat{i}}$ optimal.

With this background, we can now state Shao's main results about AIC.

Suppose the true model is not in the available model space or there is a unique true model in the available model space. Assume also a blanket condition on the true

μ_j's which prevent them from being too small (condition (2.6) of Shao[1]). Then AIC is asymptotically optimal. The penalty $\lambda = 2$ plays a special role in this, as indicated below.

Consider a general penalized likelihood rule which chooses a model minimizing the information criterion

$$-2\log f(\underset{\sim}{Y} \mid \hat{\underset{\sim}{\mu}}(i), M_i) + \lambda p_i = T(M_i), \quad \text{say.}$$

Suppose all available models are false. Then (for $\lambda = 2$ and only in this case) Shao shows that $\sigma^2 T(M_i)$ differs from $L(M_i, \hat{\underset{\sim}{\mu}}, \underset{\sim}{Y})$ by an amount which is random, but does not depend on M_i plus a negligible quantity. Thus, asymptotically, minimizing L or T will lead to the same model, i.e. the oracle will coincide with the model chosen by AIC. This proves Shao's theorem when the true model is not in the model space. The other case is also based on the above fact, but requires a little more work.

Problem with improper priors

The nonsubjective prior for μ usually proposed in Examples 1 and 2 is of the form, vide eq. (7)

$$p(\mu \mid M_2) = c.$$

This integrates to infinity and so cannot be normalized to integrate to one. Nonsubjective priors like this have been used a lot from Laplace onwards and lead to sensible estimates. For example, the posterior mean of μ given M_2 is just $\bar{X}$. It does not depend on the arbitrary c which cancels when one calculates the posterior distribution under the model. But when one evaluates M_2 by the integrated likelihood, the constant c remains; in the Laplace approximation it remains in the $O(1)$ term neglected in the BIC.

This phenomenon was first pointed out by the distinguished astrophysicist, probabilist and statistician H. Jeffreys, who is very well-known for his many seminal contributions to nonsubjective Bayesian analysis. Many nonsubjective priors, currently used, are due to him. In Examples 1 and 2, Jeffreys[20] suggested the priors $N(0, 2\sigma^2)$ and Cauchy (μ, σ) on certain logical and mathematical grounds. Calculations for Example 2 based on these priors are reported in the next section.

New cutting edge work on general problems of this type have been done by Berger and Pericchi[25,26]. They have introduced the notion of intrinsic Bayes factors (IBF) and intrinsic priors which are very similar or identical to nonsubjective priors of Jeffreys. Ghosh and Samanta[27] who provide a unified approach, show that the Cauchy prior can be viewed as a sort of intrinsic prior in Example 2.

Example 6 and more on BF, BIC and AIC

We start with Example 6 once more to motivate a few general observations that emerge from Mukhopadhyay[2] and Berger *et al.*[28].

This example which is essentially due to Stone[19] was used by him to show that the penalty of BIC is inappropriate compared with that of AIC, in the sense AIC picks a correct model more often. We note in passing that this is different from the original purpose of AIC to predict well.

It turns out that BIC itself is an inappropriate tool for this problem because the Schwarz type assumption of fixed dimensional models is not valid here. In fact if one calculates a Bayes factor $\mathrm{BF}_{1,2}$ with the p-dimensional Cauchy prior

$$p(\tilde{\mu} \mid M_2 = \frac{\Gamma((p+1)/2)}{(\pi^{(p+1)/2}\sigma^{p})}\left(1+\frac{\tilde{\mu}'\tilde{\mu}}{\sigma^{2}}\right)^{-(p+1)/2},$$

used in Zellner and Siow[29], it can be shown that $\mathrm{BF}_{1,2}$ identifies models correctly all the time, but BIC is a very poor approximation. Indeed, Berger *et al.*[28] propose a new GBIC which works well for this example and reduces to BIC if p is fixed as $n \to \infty$. In other words, the problem is not between AIC and Bayesian methods, but between AIC and an inadequate Bayesian method for this problem.

One can go a step further and ask what would be a proper Bayesian approach to prediction with squared error loss. This is quite different from the interesting predictive Bayes factors due to Aitkin[30], Geisser and Eddy[31] and Gelfand and Dey[32] and often used in the computer science literature. They have not yet been studied from the point of view of fully Bayes optimal prediction with respect to a suitable loss function.

We return to the problem posed at the beginning of the last paragraph. The solution is the Bayesian model average obtained in the following way. Let $\hat{\underset{\sim}{\mu}}^B(i)$ be the prediction or estimate of $\underset{\sim}{\mu}$ assuming M_i. Now form a weighted average using the posterior probability of a model as the weight to be attached. For many practical applications and numerical recipes in the presence of many models, see Hoeting *et al.*[33]. Asymptotic theoretical properties are not known.

In Example 6 we have only two models and it turns out that an asymptotic treatment is possible, but non-trivial. The main result is that with a nonsubjective prior like the Cauchy prior (or a normal prior with empirical Bayes estimation of parameters of the normal prior), the model average outperforms AIC by an order of magnitude!

The result throws further light on AIC. Suppose M_2 is true and $\frac{1}{p}\Sigma\mu_i^2 \to \tau^2 > 1$. Then the AIC chooses M_1 almost always, but the Bayes rule chooses M_2 almost always. But after choosing the more complex model, the Bayes rule shrinks the estimates of $\underset{\sim}{\mu}$ from their least squares values towards zero. The amount of shrinkage depends on τ^2; the smaller τ is, the more the shrinkage towards zero. AIC cannot do this because it is constrained to use least squares estimates. So to get the same sort of effect as shrinking, it chooses the very low-dimensional model M_1. The constraint of using least squares estimates leads to its poor performance compared with Bayes rules. If one constrains the Bayesian approach in the same way by requiring that once a model M_i is chosen the corresponding least squares estimates must be chosen, then the advantage of Bayes rules disappears and, in fact, it can be shown that asymptotically the Bayes rules and AIC behave in the same way.

These results have been tested in a classical example of Shibata[34] of regression with orthogonal polynomials. Instead of the signal $\underset{\sim}{\mu}$ one has a square integrable function on [0, 1]. What one observes are its values at n-equidistant points of [0, 1], corrupted by normal noise. The main difference with Example 6 is that the regression coefficients which are the analogues of μ_j's are square summable. So there is less and less additional information as the number of observations increase. Bayes rules still outperform AIC but not by an order of magnitude. Once again the advantage disappears if the Bayes rule is constrained to use least squares estimates.

It is interesting and ironic that AIC has a Bayesian justification for high dimensions, if least squares estimates have to be used, while the BIC has no Bayesian credentials. However, while these results have been verified in a number of non-trivial examples by asymptotics or simulations, no general theory covering high-dimensional problems is still available. Development of such a theory is a hard mathematical problem. In the short run, extensive and well-chosen simulations may be more successful.

The corrected AIC (AIC$_c$) has not been studied as carefully as AIC.

For alternative points of view on empirical Bayes model selection see George and Foster[35].

Calculations for the examples

We present below analysis of the data for some of the examples presented earlier in the article.

Example 1 and 2: We analyse the data on deflection of light presented in Example 2 to examine whether Einstein's prediction of $\mu = 1.75$ is supported by the data. The four independent observations are assumed to have an $N(\mu, \sigma^2)$ distribution. Here μ is the parameter of interest and σ^2 represents variation in the error of measurements.

Even though σ^2 is not known, for the sake of illustration, we first take it to be known and equal to the sample variance s^2. Note that this is only a simplification of the problem and reduces the problem to that of Example 1. We consider a transformation $x' = (x - 1.75)/s$ of the original data x, so that the transformed observations may be assumed to have an $N(\mu, 1)$ distribution with $\mu = 0$ under M_1. If we use an $N(0, 2)$ prior as suggested by Jeffreys, the Bayes factor BF$_{1,2}$ of M_1 relative to M_2 (defined in eq. (6)) is calculated as 3.0. For M_1 the common value of AIC and BIC is obtained as 11.35, while for M_2, AIC is 13.35 and BIC is 12.74. Thus the calculations with the given data lend some support to Einstein's prediction. However, the evidence in the data is not very strong. This particular experiment has not been repeated because of unavoidable experimental errors. There are now better confirmations of Einstein's theory, vide Gardner[16].

If σ^2 is not assumed to be known, as is the case, one has to specify appropriate priors $\pi_1(\sigma)$ and $\pi_2(\mu, \sigma)$ under M_1 and M_2, respectively, to calculate the Bayes factor BF$_{1,2}$. The conventional priors suggested by Jeffreys are

$$\pi_1(\sigma) = \frac{1}{\sigma}, \pi_2(\mu, \sigma) = \frac{1}{\sigma} \cdot \frac{1}{\pi\sigma(1 + (\mu - \mu_0)^2/\sigma^2)},$$

where μ_0 is the value of μ specified under M_1 which is 1.75 in our case. With these priors BF$_{1,2}$ for our example turns out to be 2.98. The values of AIC for M_1 and M_2 are calculated as 6.01 and 8.01 respectively, while the respective values of BIC are 5.39 and 6.78.

Example 3: Hald's regression data have been analysed by Berger and Pericchi[36] and Burnham and Anderson[7] to select a linear regression model. Burnham and Anderson[7] used the AIC and its variant, AIC$_c$ (adjustment for small sample size), while Berger and Pericchi[36] reported the BIC, the intrinsic Bayes factors (IBF) and the Bayes factor based on the conventional prior of Zellner and Siow[29] which is a generalization of the conventional Jeffreys prior used in Example 2. There are four regressors and hence $2^4 - 1 = 15$ possible models involving at least one of the regressor variables. We represent a model by the labels of the regressor variables chosen. For example, a model that uses only the regressor variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is denoted by {12}. Table 3 presents the values of AIC, AIC$_c$, BIC, Arithmetic IBF (AIBF) of Berger and Pericchi[25] and Bayes factor with Zellner–Siow prior (ZSBF). The Bayes factors (AIBF and ZSBF) presented are those of the full model {1234} relative to all possible models. Calculations of AIBF and ZSBF are obtained from Berger and Pericchi[36]. Since these criteria are on a relative scale we report only their differences (Δ) from the respective minimum value over the models (e.g. Δ(AIC) = AIC – min AIC), so that the model with $\Delta = 0$ is to be selected. Models are ordered in terms of the AIC$_c$ differences. Use of AIC$_c$ rather than AIC is recommended for the data as the sample size is relatively small compared to the number of parameters (Burnham and Anderson[7]). Interestingly, all the criteria (except AIC) select the same model {12}. AIC chooses {124}, which is also a plausible model, vide Burnham and Anderson[7].

Table 3. Δ values for AIC, AIC$_c$ BIC, 2log(AIBF) and 2log(ZSBF) together with number of estimable parameters (k) for different models for Hald's regression data of Table 1

Model	k	AIC (Δ)	AIC$_c$ (Δ)	BIC (Δ)	2log(AIBF) (Δ)	2log(ZSBF) (Δ)
{12}	4	0.45	0.00	0.00	0.00	0.00
{124}	5	0.00	3.13	2.73	0.74	0.29
{123}	5	0.04	3.16	2.65	0.95	0.29
{14}	4	3.77	3.32	3.46	1.88	1.53
{134}	5	0.75	3.88	3.40	1.09	0.65
{234}	5	5.60	8.73	8.31	3.79	2.90
{1234}	6	1.97	10.52	5.06	3.43	2.69
{34}	4	14.88	14.43	14.80	8.57	6.63
{23}	4	26.06	25.62	25.82	14.18	11.71
{4}	3	33.88	31.10	29.60	16.92	16.68
{2}	3	34.20	31.42	29.78	18.54	16.83
{24}	4	35.66	35.21	34.42	19.29	16.15
{1}	3	38.55	35.77	32.18	20.10	18.79
{13}	4	40.14	39.70	36.84	21.46	18.29
{3}	3	44.09	41.31	37.90	23.50	21.39

Table 4. Δ values for AIC and – 2log(Int L) together with number of estimable parameters (k) for different models for the starling data of Table 2

Model	k	AIC (Δ)	– 2log(Int L) (Δ)
$M_{0,0}$	15	45.57	29.97
$M_{1,0}$	16	34.46	23.72
$M_{1,1}$	17	27.99	20.15
$M_{2,1}$	18	12.78	7.28
$M_{2,2}$	19	3.59	0.44
$M_{3,2}$	20	0.60	0.00
$M_{3,3}$	21	0.00	1.38
$M_{4,3}$	22	1.17	4.49
$M_{4,4}$	23	2.90	8.81

Example 5: Burnham and Anderson[7] compared the possible models on the basis of the AIC and selected the model $M_{3,3}$ (M_{4p} in their notation). Table 4 gives the values of the differences Δ = AIC – min(AIC) for the first 9 models $M_{0,0}, \ldots, M_{4,4}$ as obtained by them. We have also calculated the integrated likelihoods for these models assuming independent $U(0, 1)$ priors for the parameters. We have used the *importance sampling* method of numerical integration (see ref. 37) to calculate the integrated likelihoods for the models. Note that the parameter spaces are of high dimensions and we indeed have sampled from an appropriate neighbourhood of the maximum likelihood estimators (MLE) of the parameters. We first calculated the integral over the region MLE ± d_0, where the choice of d_0 was guided by an idea about the standard errors of the estimates. We then calculated the integrals over MLE ± d_i, $i = 1, 2, \ldots$ for some suitable $d_0 < d_1 < d_2 < \ldots$ until the integrals converge. Here we have taken $d_0 = 0.1$, $d_1 = 0.15$ and $d_2 = 0.2$ and sampled 500–1000 million times for each of the models.

The values reported in column 4 of Table 4 are differences (Δ) for logarithm of the integrated likelihoods multiplied by – 2 (i.e. – 2log(Int L)). The Bayes factor of a model relative to another may be obtained as the ratio of the corresponding integrated likelihoods. Note that on the basis of the integrated likelihood we select the model $M_{3,2}$, a model with less number of parameters than that selected by the use of AIC.

Concluding remarks

We have tried to motivate BIC and AIC and present new facts which have either been published very recently or are still to appear. We also show these in action in a few interesting examples. In actual problems, the difference between BIC and AIC or the Bayes factor is not as much as popularly believed. This is because the complexity of the models used depends on the size of the available data. In these examples, models get complex as data increase so that the conflict in the two criteria, for low-dimensional models tested on large data sets, does not appear here. There is some theoretical support for this in Mukhopadhyay[2].

One striking fact is that Bayes rules and AIC, properly interpreted, may not be as different as they are often assumed to be. Another interesting fact is that the common perception about penalty being much more severe in Bayes rules is not correct. Not only are the penalties much more similar than currently perceived, Bayes rules may select more complex models than AIC in high-dimensional problems.

Bayesian model averages is an extremely powerful new tool. It seems to do better than AIC, by an order of magnitude in some cases. But in most real-life problems, this comes with a price of heavy computations.

Finally, the effect of the preferred loss function, zero-one or squared error, has to be taken into account. One has to decide, in Stone's words, whether it is going to be hard or soft science, i.e. whether one wants to know the truth or predict well. In the first case, one should use a Bayes factor or a good approximation. The BIC is not a good approximation if the dimension is large. In the second case, one should use the AIC or the model average or a model that provides the closest prediction to model average. There is some evidence that the last two will do better than AIC.

1. Shao, J., *Stat. Sin.*, 1997, **7**, 221–264.
2. Mukhopadhyay, N., Ph D thesis submitted, Purdue University, 2000.
3. Bartholomew, D. J. and Knott, M., *Latent Variable Models and Factor Analysis*, Arnold, London, 2nd edn, 1999.
4. Agresti, A., *Categorical Data Analysis*, John Wiley and Sons, New York, 1990.
5. Lindsay, J. K., *Applying Generalized Linear Models*, Springer–Verlag, New York, 1997.
6. McCullagh, P. and Nelder, J. A., *Generalized Linear Models*, Chapman and Hall, New York, 1989, 2nd edn.
7. Burnham, K. P. and Anderson, D. R., *Model Selection and Inference – A Practical Information – Theoretic Approach*, Springer-Verlag, New York, 1998.
8. Green, P. J., *Biometrika*, 1995, **82**, 711–732.
9. Grenander, U. and Miller, M., *J. R. Stat. Soc. Ser. B*, 1994, **56**, 549–603.
10. MacKay, D. J. C., Ph D thesis, California Institute of Technology, 1991.
11. Mackay, D. J. C., *Network Comput. Neural Systems*, 1995, **6**, 469–505.
12. Sundararajan, S., Ph D thesis submitted, Indian Institute of Science, Bangalore, 1999.
13. Li, M. and Vitanyi, P., *An Introduction to Kolmogorov Complexity and its Applications*, 1993.
14. Rissanen, J., *Ann. Stat.*, 1986, **14**, 1080–1100.
15. Rissanen, J., *J. R. Stat. Soc. Ser. B*, 1987, **49**, 223–239.
16. Gardner, M., *Relativity Simply Explained*, Dover, Mineola, New York, 1997.
17. Woods, H., Steinour, H. H. and Starke, H. R., *Ind. Engg. Chem.*, 1932, **24**, 1207–1214.
18. Stromborg, K. L., Grue, C. E., Nochols, J. D., Hepp, G. R., Hines, J. E. and Bourne, H. C., *Ecology*, 1988, **69**, 590–601.
19. Stone, M., *J. R. Stat. Soc. Ser. B*, 1979, **41**, 276–278.
20. Jeffreys H., *Theory of Probability*, Oxford University Press, London, 1961.
21. Schwarz, G., *Ann. Stat.*, 1978, **6**, 461–464.
22. Akaike, H., in Proceedings of the Second International Symposium on Information Theory (eds Petrov, B. N. and Czaki, F.), Akad. Kiado, Budapest, 1973, pp. 267–281.
23. Akaike, H., *Ann. Inst. Stat. Math.*, 1978, **30**, 9–14.
24. Stone, M., *Stat. Sin.*, 1997, **7**, 252–254.
25. Berger, J. and Pericchi, L., *J. Am. Stat. Assoc.*, 1996, **91**, 109–122.
26. Berger, J. O. and Pericchi, L. R., ISDS Discussion Paper, Duke University, 2000.
27. Ghosh, J. K. and Samanta, T., *J. Statist. Plann. Infer.*, 2001 (to appear).
28. Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N., ISDS Discussion Paper, Duke University, 2000.
29. Zellner, A. and Siow, in *Bayesian Statistics 1* (eds Bernardo, J. M. *et al.*), Valencia University Press, Valencia, 1980, pp. 585–603.
30. Aitkin, M., *J. R. Stat. Soc. Ser. B*, 1991, **53**, 111–142.
31. Geisser, S. and Eddy, W. F., *J. Am. Stat. Assoc.*, 1979, **74**, 153–160.
32. Gelfand, A. and Dey, D., *J. R. Stat. Soc. Ser. B*, 1994, **56**, 501–514.
33. Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. *Stat. Sci.*, 1999, **14**, 382–417.
34. Shibata, R., *Ann. Inst. Stat. Math.*, 1983, **35**, 415–423.
35. George, E. I. and Foster, D. P., Technical Report, The University of Texas at Austin and the University of Pennsylvania, 2000.
36. Berger, J. and Pericchi, L., in *Bayesian Statistics* (eds Bernardo, J. M. *et al.*), Oxford University Press, London, 1995, vol. V, pp. 23–42.
37. Berger, J. O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, 1985, 2nd edn.

ACKNOWLEDGEMENTS. Results discussed in the section 'Example 6 and more on BF, BIC and AIC' are based on joint work by James O. Berger, Nitai Mukhopadhyay and one of the authors (JKG). However, only we are responsible for any philosophical conclusions drawn from the mathematical facts. We thank Ms Kajal Dihidar for some help in computation.