The Curve Fitting Problem: A Bayesian Rejoinder

Prasanta S. Bandyopadhyay,[1] Montana State University

Robert J. Boik, Montana State University

September 17, 1998

[1]Portions of this paper will be presented by the senior author at the APA, Central Division, 1998

The Curve Fitting Problem: A Bayesian Rejoinder

Abstract

In the curve fitting problem two conflicting desiderata, simplicity and goodness-of-fit pull in opposite directions. To solve this problem, two proposals, the first one based on Bayes' theorem criterion (BTC) and the second one advocated by Forster and Sober based on Akaike's Information Criterion (AIC) are discussed. We show that AIC, which is frequentist in spirit, is logically equivalent to BTC, provided that a suitable choice of priors is made. We evaluate the charges against Bayesianism and contend that AIC approach has shortcomings. We also discuss the relationship between Schwarz's Bayesian Information Criterion and BTC. [Word count 93]

Overview

In the curve fitting problem, two conflicting desiderata, simplicity and goodness-of-fit, pull in opposite directions. Simplicity forces us to choose straight lines over non-linear equations, whereas goodness-of-fit forces us to choose the latter over the former. This article discusses two proposals that attempt to strike an optimal balance between these two conflicting desiderata. A Bayesian solution to the curve fitting problem can be obtained by applying Bayes' theorem. The Bayesian solution is called the Bayes' Theorem Criterion (BTC). Malcolm Forster and Elliot Sober, in contrast, propose Akaike's Information Criterion (AIC) which is frequentist in spirit. The purpose of this article is threefold. First, we address some of the objections to the Bayesian approach raised by Forster and Sober. Second, we describe some limitations in the the implementation of the approach based on AIC. Finally, we show that AIC is in fact logically equivalent to BTC with a suitable choice of priors. The underlying theme of this paper is to illuminate the Bayesian/non-Bayesian debate in philosophy of science.

1 The Curve Fitting Problem: BTC

Those who live in colder climates have to face this problem at one time or another especially if they heat their home with natural gas. Sue is one such woman. The amount of gas required to heat her home depends on the outdoor temperature — the colder the weather, the more gas will be consumed. As long as the family's habits, the insulation of the house, and other such factors don't change, Sue should be able to predict gas consumption from the outdoor temperature. The usual need for heating is measured in degree days. One heating degree day is accumulated for each degree the average daily temperature falls below 65° Fahrenheit. An average temperature of 20° F., for example, corresponds to 45 degree days. Table 1 presents historical data for one season. In Table 1, the explanatory variable, x, is heating degree days for the month, and the response variable, Y, is gas consumption per day in units of 100 cubic feet.

Table 1

	Month								
Variable	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
x	15.6	26.8	37.4	36.4	35.5	18.6	15.3	7.9	0.0
Y	5.2	6.1	8.7	8.5	8.8	4.9	4.5	2.5	1.1

Sue's goal is to use these data to predict gas consumption at different temperatures. To do this she wants to find a relationship, if any between, x and y. The most common technique for fitting a line to data is known as the *method of least squares* and the line drawn that runs through the scatter plot is known as the *regression line*. It is assumed that Y_i is related to x_i through the linear model

$$Y_i = \alpha_0 + \sum_{j=1}^{k} \alpha_j x_i^j + \varepsilon_i, \text{ for } i = 1, \ldots, n,$$

where n is the sample size; α_j, $j = 0, \ldots, k$ are unknown regression coefficients, k is the order of the polynomial model, and ε_i is random error. The error terms, ε_i, $i = 1, \ldots, n$ are assumed to be independently distributed as Gaussian random variables with mean zero and variance σ^2.

Sue wants to know what her gas consumption will be for a month this year that has $x = 15.3$ degree days. In April of the historical data set, 15.3 heating degree days corresponded to 4.5 hundred cubic feet of gas. We shall forecast from three different regression line equations how much gas the house will require at 15.3 degree days. Each regression equation corresponds to a hypothesis, H_1, H_2 or H_3 in a domain in which each is mutually exclusive of the others. The hypotheses are obtained by assuming that the order of the polynomial regression model is 1, 2, or 3. That is, $\mathrm{H}_1\colon \mathrm{E}(Y|x) = \alpha_0 + \alpha_1 x$; $\mathrm{H}_2\colon \mathrm{E}(Y|x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$; and $\mathrm{H}_3\colon \mathrm{E}(Y|x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$. Here, $\mathrm{E}(Y|x)$ is the conditional expectation of Y given x. To say that these hypotheses are mutually exclusive is to say that the coefficient of x^k under H_k is non-zero.

The least squares forecast under H_1 is $\widehat{Y} = 1.22 + .20 \times 15.3 = 4.33$ hundred cubic feet. Under H_2, the prediction is $\widehat{Y} = 1.09 + .22 \times 15.3 - 0.0005 \times 15.3^2 = 4.39$ hundred cubic feet per day. This is closer to the historical value of 4.5 than that based on H_1. If we use H_3, then we will find that the prediction, $\widehat{Y} = 4.45$ is even closer to the historical value. In general, as the order of the polynomial regression model increases, the goodness-of-fit of the model to the observed data increases. A standard measure of the goodness-of-fit is the likelihood function evaluated at the parameter values which give a maximum value of the function. Denote the value of the maximized likelihood function Under H_k by $\widehat{L}_k$. That is

$$\widehat{L}_k = \max_{\sigma^2;\alpha_0,\ldots,\alpha_k} L_k(\sigma^2; \alpha_0, \ldots, \alpha_k | \mathrm{H}_k; Y_1, \ldots, Y_n), \tag{1}$$

where L_k is the likelihood function.

As the order of the polynomial model increases, the maximized likelihood increases. That is, $\widehat{L}_k$ is an increasing function of k. Thus, it might appear that as the order of the model increases, the predictions will become more accurate. The fact of the matter, however, is otherwise. A model having too large of an order will over-fit the data. Predictions of future data from such a model will, in general, have larger errors than will predictions from a model with a smaller, but sufficient, number of parameters.

Bandyopadhyay et al. (1996) employed Bayes Theorem and showed that if one adopts certain non-informative priors on σ^2 and α_j for $j = 1, \ldots, k$, then the posterior probability of H_k, given the data, is proportional to the maximized likelihood function, $\widehat{L}_k$, multiplied by the prior probability of H_k. That is,

$$\Pr(\mathrm{H}_k|\text{data}) \propto \widehat{L}_k \times \Pr(\mathrm{H}_k), \tag{2}$$

where $\widehat{L}_k$ is the maximized likelihood function given in (**??**) and $\Pr(\mathrm{H}_k)$ is the prior probability of H_k reflecting the agent's beliefs before any evidence is known. The right-hand-side of (**??**) is one application of BTC. An investigator using this application of BTC would compute the posterior probability of each hypothesis under consideration and if one is to be selected, it would be the one with the highest posterior probability.

To obtain the posterior probability in (**??**), Bandyopadhyay et al. adopted a Gaussian prior on the vector of regression coefficients. The prior variance of the regression coefficients was assumed to be of order of magnitude $\frac{1}{n}$. If the prior on the regression coefficients is redefined so that it does not depend on n, then the posterior probability of H_k, given the data, can be shown to be

$$\Pr(\mathrm{H}_k|\text{data}) \propto \widehat{L}_k \times \Pr(\mathrm{H}_k) \times n^{-k/2}. \tag{3}$$

The right-hand-side of (**??**) is a second application of BTC and is equivalent to Schwarz's (1978) Bayesian Information Criterion (BIC), provided that $\Pr(\mathrm{H}_k)$ is not a function of n.

The posterior distributions in (**??**) and (**??**) also can be obtained by adopting other non-informative priors. In particular, the invariant prior of Jeffreys (1961) leads to (**??**) if information is computed for a sample of size n and to (**??**) if information is computed for a sample of size 1. In fact, as shown by Schwarz (1978), the posterior distribution in (**??**) is obtained (to a first order approximation) by adopting any prior distribution on σ^2 and $\alpha_0, \ldots, \alpha_k$ from a large class of informative and non-informative priors. One necessary characteristic of the priors in this class is that they do not depend on the sample size, n.

In the application of BTC, two factors, formal and non-formal, determine the prior probability of the hypothesis, $\Pr(\mathrm{H}_k)$. The formal factor is paucity of parameters and this factor orders hypotheses with respect to simplicity. Recall the three hypotheses, H_1, H_2, and H_3, where the order of the polynomial model is denoted by the subscript. A hypothesis gets a higher probability than a competitor, *ceteris paribus*, if it has fewer parameters. That is, $\Pr(\mathrm{H}_1) > \Pr(\mathrm{H}_2) > \Pr(\mathrm{H}_3)$.

Restricting the prior probabilities of the hypotheses to satisfy $\Pr(\mathrm{H}_i) > \Pr(\mathrm{H}_j)$ whenever $i < j$ is not sufficient to determine the values of these probabilities. Non-formal factors, especially epistemological and pragmatic factors, play a key role in arriving at the specific values of the prior probabilities. Several approaches have been suggested for how investigators might use historical data, expert opinion, and other epistemological and pragmatic factors to assign specific prior probabilities (e.g., Berger, 1985). For illustrative purposes, our discussion will focus on just three of many possible priors. These three priors are summarized in Table 2.

Table 2

Prior	Equation
$\Pr_1(\mathrm{H}_k)$	$2^{-k}; \quad k = 1, 2, \ldots$
$\Pr_2(\mathrm{H}_k)$	$(e-1)e^{-k}; \quad k = 1, 2, \ldots$
$\Pr_3(\mathrm{H}_k)$	$(\sqrt{n}-1)n^{-k/2}; \quad k = 1, 2, \ldots$

In Section 4, we will use the posterior probabilities in equations (**??**) and (**??**) along with the priors in Table 2 to choose the polynomial order for the Sue data in Table 1. First, however, we will describe several charges that have leveled against the Bayesian approach.

2 Charges Against Bayesianism

There are allegedly several problems with the Bayesian approach. First, some charge that Bayesianism leads to subjectivism. Bayesianism is suspect, it is claimed, because subjectivism vitiates a scientific approach to making inferences.

Bayesianism derives its name from its heavy use and interpretation of Bayes' theorem. This theorem relies on the notion of prior probability. An agent's prior probability for a hypothesis represents that agent's belief in the hypothesis prior to conducting the planned investigation. For the same hypothesis, another agent could have a different prior probability. An agent's prior belief in a specific hypothesis can be any probability ranging from 0 to 1, exclusively. Because Bayesianism allows two agents to start with non-extreme divergent priors, provided their assignments of priors are consistent with the probability calculus, Bayesians are sometimes branded as subjectivists. In our Bayesian account, an agent might choose the prior probability $\Pr_1(H_k)$ in Table 2. According to the first objection, this assignment of probabilities is subjective because it is arbitrary and devoid of any justification.

A second charge against Bayesianism is that the goal of the BTC approach is flawed. In the Sue example, as Bayesians we consider three hypotheses as possible candidates for predicting future data. We choose the hypothesis H_k if the probability of H_k, given the data, is highest. Forster (1997) objects that working scientists are not interested in probability of truth. Therefore, attempting to select the hypothesis with the highest posterior probability of being correct is a fruitless exercise. Instead, working scientists are interested in predictive accuracy. According to this objection, no working scientist acts like a Bayesian.

Third, Forster and Sober (1994) charge that the approach based on BTC is flawed because it fails to deal with the importance of *families* of curves. The charge states that to speak of the probability of the best fitting curve being true is nonsense because no amount of data will produce a curve that is exactly true.

To escape these problems Forster and Sober suggest that an approach based on Akaike's Information be used, rather than an approach based on BTC. In practice, Akaike's approach is implemented by means of Akaike's Information Criterion (AIC).

3 The Curve Fitting Problem: AIC

In the curve fitting problem, according to Forster and Sober, we would like to choose the curve that is closest to the true curve. Their desire is to measure the closeness of a family of curves to truth without invoking prior probability. To accomplish this goal, they suggest that AIC be used. The theory underlying AIC assumes that there is a true distribution of the observable random variables. Call this distribution $f(\mathbf{Z}|\boldsymbol{\theta})$ (or f) where $\mathbf{Z}$ is a vector of observable random variables (i.e., future data) and $\boldsymbol{\theta}$ is a vector of unknown parameters (e.g., $\widehat{\sigma}^2$ and $\widehat{\alpha}_0, \ldots, \widehat{\alpha}_k$). Under H_k, an approximation to the unknown distribution, f, is desired. In the Akaike approach one approximates f in two steps. First, a sample of data, say $\mathbf{Y}$, is observed and $\widehat{\boldsymbol{\theta}}_k$, the maximizer of $f_k(\mathbf{Y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is computed. Second, the unknown density, f, is approximated by $\widehat{f}_k = \widehat{f}_k(\mathbf{Z}|\widehat{\boldsymbol{\theta}}_k)$

To make predictions with optimal accuracy, one would use f if the density was known. In the absence of such knowledge, $\widehat{f}_k$ is used, where the value of k is chosen to minimize the distance between f and $\widehat{f}_k$. Akaike (1973) argued that a useful measure of the distance between f and $\widehat{f}_k$ is twice the expected Kullback-Leibler information. That is, Akaike's measure of the distance between f and $\widehat{f}_k$ is

$$D(f, \widehat{f}_k) = \frac{2}{n} \mathrm{E}\left[\int \ln\left(\frac{f}{\widehat{f}_k}\right) f \, d\mathbf{Z}\right],$$

where the expectation is taken with respect to the distribution of the estimator $\widehat{\boldsymbol{\theta}}_k$. By taking the latter expectation, Akaike obtained a measure of the distance between the true curve and a family of curves rather than the distance between the true curve and a specific member of the family.

Choosing the family of curves which minimizes $D(f, \widehat{f}_k)$ is equivalent to choosing the family of curves which maximizes the quantity

$$A(f, \widehat{f}_k) = \frac{2}{n} \mathrm{E}\left[\int \ln\left(\widehat{f}_k\right) f \, d\mathbf{Z}\right] \tag{4}$$

Forster and Sober (1994) refer to $A(f, \widehat{f}_k)$ as the predictive accuracy of the k^{th} family of curves. The negative of predictive accuracy also is known as

Kullback-Leibler divergence.

Akaike showed that, under certain conditions,

$$\widehat{A}(f, \widehat{f}_k) = \frac{2}{n} \ln\left(\widehat{L}_k\right) - \frac{2}{n} \mathrm{Dim}(\mathrm{H}_k) + \text{ constants}$$

is a consistent estimator $A(f, \widehat{f}_k)$, where $\widehat{L}_k$ is the maximized likelihood function given in equation (**??**) and $\mathrm{Dim}(\mathrm{H}_k)$ is the number of adjustable parameters estimated under H_k. For the polynomial models under consideration, $\mathrm{Dim}(\mathrm{H}_k) = k + 2$. To maximize predictive accuracy, one chooses the hypothesis which maximizes

$$\mathrm{AIC} \ = \ln\left(\widehat{L}_k\right) - k. \tag{5}$$

4 AIC is Equivalent to BTC With a Suitable Choice of Priors

Consider the two criteria BTC and AIC. BTC chooses the hypothesis that maximizes the posterior probability whereas AIC chooses the hypothesis that maximizes predictive accuracy. The two approaches are summarized in Table 3.

Table 3

Approach	General Criterion	Specific Prior	Specific Criterion (Log Scale)	Applied to Sue Data H_1	H_2	H_3
BTC_1	$\widehat{L}_k \Pr(\mathrm{H}_k)$	$\Pr_1(\mathrm{H}_k)$	$\ln(\widehat{L}_k) - k\ln(2)$	-4.75	-5.30	-5.55
		$\Pr_2(\mathrm{H}_k)$	$\ln(\widehat{L}_k) - k$	-5.06	-5.92	-6.47
		$\Pr_3(\mathrm{H}_k)$	$\ln(\widehat{L}_k) - \frac{k}{2}\ln(n)$	-5.15	-6.11	-6.77
BTC_2	$\widehat{L}_k \Pr(\mathrm{H}_k) n^{-k/2}$	$\Pr_1(\mathrm{H}_k)$	$\ln(\widehat{L}_k) - \frac{k}{2}\ln(n)$	-5.15	-6.11	-6.77
		$\Pr_2(\mathrm{H}_k)$	$\ln(\widehat{L}_k) - \frac{k}{2}\ln(n)$	-5.15	-6.11	-6.77
AIC	$\widehat{L}_k e^{-k}$	—	$\ln(\widehat{L}_k) - k$	-5.06	-5.92	-6.47

For each approach in Table 3, the hypothesis selected is the one which maximizes the criterion. The approaches labeled BTC_1 and BTC_2 correspond to the posterior probabilities in equations (??) and (??). The hypothesis which is selected using BTC_1 potentially depends on the specific prior adopted. Hypothesis selection using BTC_2 also depends on the specific prior adopted, but only minimally. If $\Pr(\mathrm{H}_k)$ is not a function of n, then $\ln[\Pr(\mathrm{H}_k)] - \frac{k}{2}\ln(n)$ is dominated by the term $-\frac{k}{2}\ln(n)$. Accordingly, for moderate to large samples, the selection can be based on the maximized log likelihood function minus the penalty function $\frac{k}{2}\ln(n)$. The approach based on AIC is not Bayesian and, therefore, the hypothesis selected by AIC does not depend on any prior probabilities.

Column 4 of Table 3 displays the specific criterion obtained when specific priors are adopted. Clearly AIC is equivalent to BTC_1 when the prior $\Pr_2(\mathrm{H}_k)$ is adopted. Equivalence under other priors also can be demonstrated. For example, Akaike (1978) adopted an informative prior on the parameters (σ^2, $\alpha_0, \alpha_2, \ldots$) but weighted all families equally. Under his specific conditions, Akaike showed that BTC is equivalent to AIC.

Note that all criteria can be written as the maximized log likelihood function minus a penalty function. The larger the penalty function, the more weight the criterion gives to simplicity. If $n \geq 8$, the three penalty functions in Table 3 can be ordered as follows:

$$k\ln(2) < k < \frac{k}{2}\ln(n).$$

Accordingly, BTC can give more, equal, or less weight to simplicity than does AIC. The last three columns in Table 3 apply the various criteria to the Sue data. For this data set, all criteria select H_1 as the best trade-off between goodness-of-fit and simplicity.

5 A Bayesian Response

5.1 Subjectivity

Our response to the charge that the use of priors smacks of subjectivism and arbitrariness has several parts. First, we contend that any assignment of priors must satisfy four constraints.

1. It must satisfy the probability calculus.
2. It must order hypotheses with respect to simplicity.
3. It must be buttressed by epistemological/pragmatic considerations.
4. It must follow a rule.

In most applications, the priors on a finite set of hypotheses can be ordered in an objective manner. We argued in Section 1 and elsewhere (Bandyopadhyay et al, 1996) that the prior probability of a theory depends on simplicity as well as other factors. Paucity of adjustable parameters provides a formal index of simplicity. Lower order polynomial curves get higher prior probability than higher order polynomial curves.

In some applications, paucity of parameters may not suffice to order hypotheses with respect to simplicity. The ordering of prior probabilities sometimes depends on epistemological/background information. For instance, scientists routinely ignore solutions to equations which are not sensible in a given situation (e.g., negative values for quantities which can only plausibly be positive). We reject these values because on the basis of our background knowledge we know that they make no sense.

Whether an equation is mathematically tractable, or whether that equation is easy to handle plays a vital role in theory choice, hence in the assignment of prior probabilities to theories. Working scientists take resort to this reason frequently. As an illustration, consider an example from Weinberg (1992). Each of the following is an infinite series: (a) $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ and (b) $1 + 2 + 3 + \cdots$. Each series is infinite but, according to Weinberg, "one is less infinite than the other, in the sense that it

takes less effort to figure out what to do about it" (pp. 112). Given our selection criterion of simplicity, we will take the latter to be simpler than the former because we take it to be easy to comprehend the former than the latter.

As illustrated in Table 2, ordering priors with respect to simplicity does not dictate the specific form of the prior. That is, an agent can still choose any of the three priors or some other prior that orders hypotheses with respect to simplicity. Here, an agent can take one of two approaches. First, the agent can employ pragmatic and/or epistemological factors to choose the specific form of the prior. An agent adopting this approach is responsible for justifying his/her choice. If the prior can not be justified, then the criticism of subjectivity is well founded.

Second, in some cases the agent may be able to show that for priors chosen from a well defined class, the results are invariant with respect to the specific choice of a prior from that class. In this case, the agent need only defend the class of priors and not any specific choice. Schwarz's BIC can be defended in this manner. Priors chosen ¿from a large class all yield the same criterion to a first order approximation. For moderate to large sample sizes, these priors play a negligible role.

5.2 Probability of Truth

The objection says that a Bayesian approach with its focus on the probability of a particular hypothesis being true fails to agree with practice of science. According to the objection, scientists' only concern is with predictive accuracy. We have not done a survey among working scientists to examine the above claim because we think, as Bayesians, that it would be very expensive to do so.

We believe that all scientists do not have the same goals. Certainly, many scientists are interested in predictive accuracy and we have no qualms with this. Nonetheless, the goal of the BTC approach is reasonable whenever the scientist believes that the true model is a member of one of the approximating families under consideration. In designed experiments, this is usually the case. In a designed experiment, potentially confounding variables are controlled by means of blocking or randomization. The true model, then, is a function of a subset of the known

explanatory variables. The specific subset of variables which yields the true model, however, is not known. The investigator may want to determine which subset of the explanatory variables yields the true model.

5.3 Importance of Families

The charge that the BTC approach fails to deal with the importance of families is based on a misconception about BTC. In the Sue example, we desire to know the polynomial order of the true model. For each k, there is a family of polynomial regression models which differ in the values of their regression coefficients.. The hypothesis H_k states that the true model is *some* member of this family. The hypothesis H_k does not refer to any specific k^{th} order polynomial. The goal of the BTC analysis is to compute the posterior probability that the true model is a member of the k^{th} family of polynomials. The BTC approach does not attempt to compute the posterior probability of some specific member of the family being true.

Forster and Sober (1994) argue that Bayesians may try to compare likelihoods of *families* of curves, but they are unsuccessful. First they argue that if a strictly information-less prior on the unknown parameters is adopted, then the probability density of the data averaged over all values of the unknown parameters must be zero. This is true, but irrelevant. The posterior probability of the family is a ratio of probability densities. The numerator and denominator may each go to zero as the information in the prior goes to zero, but the ratio of densities does not go to zero. Second, they argue that if certain almost information-less priors are adopted, then the BTC approach suffers because it is not invariant under reparameterization. In response, we ask them to justify their choice of priors almost-information-less priors.[1]

[1]Also, we remind them not to neglect the Jacobian when performing transformations. See footnote number 37 in Forster and Sober (1994).

6 Problems with AIC

6.1 Consistency

Consider the situation where the true family of curves is among the set approximating families. The true family is defined as the lowest order curve that correctly models the data. The AIC approach has been criticized in this situation because as sample size increases to infinity, AIC will not necessarily select the true family. Forster (1997) has addressed this charge. We do not take issue with his response. The goal of AIC is to choose a family of models that has highest predictive accuracy and this is not the same as choosing the family that has the highest posterior probability. Nonetheless, there is an issue of consistency for which AIC does not fare well.

Forster (1997) claims that AIC (when suitably scaled) is a consistent estimator of predictive accuracy. This is true in the special case of regression models where σ^2 is a known constant. In addition, if one is willing to assume that the approximating family is identical to the true family of models, then AIC is a consistent estimator of predictive accuracy. Forster's claim, however, is not true, in general. If the approximating family misspecifies the true family, then AIC no longer is consistent. For regression models, consistent estimators of predictive accuracy have been proposed (Sawa, 1978), but these estimators are of little use in more general settings. Linhart and Zucchini (1986) give a consistent estimator of predictive accuracy for the general setting, but it requires that the true family of models be specified. In practice, it may not be possible to specify the true family.

To be fair, Forster (1997) makes it clear that he is advocating the Akaike approach, not necessarily AIC. That is, Forster's goal is to maximize predictive accuracy, but he does not claim that AIC is the optimal way to reach this goal. Nonetheless, the problem remains that when using the Akaike approach, one cannot, in general, consistently estimate the quantity of interest.

6.2 Maximizing Predictive Accuracy

Suppose that one is interested in maximizing predictive accuracy. Again, we have no objection to this goal. We do wonder, though, why $\widehat{f}_k(\mathbf{Z}|\widehat{\boldsymbol{\theta}}_k)$ is used as the approximating density (see Section 3). Aitchison (1975) showed that if a prior distribution is adopted for the unknown parameter vector $\boldsymbol{\theta}$, then the density that maximizes predictive accuracy is $f_k(\mathbf{Z}|\mathbf{Y})$ rather than $\widehat{f}_k(\mathbf{Z}|\widehat{\boldsymbol{\theta}}_k)$. That is, the best predictive density for future data $\mathbf{Z}$ is the conditional density for $\mathbf{Z}$ given the observed data $\mathbf{Y}$. This density is called the posterior predictive density.

Furthermore, the posterior predictive density depends only negligibly on the prior distribution adopted for $\boldsymbol{\theta}$. For moderate to large samples, it does not matter much which prior is adopted. For example, when selecting a family of regression models, the posterior predictive density, to order $\frac{1}{\sqrt{n}}$, is a multivariate t. In general, the gain in predictive accuracy (equation **??**) from using $f_k(\mathbf{Z}|\mathbf{Y})$ rather than $\widehat{f}_k(\mathbf{Z}|\widehat{\boldsymbol{\theta}}_k)$ is approximately $\frac{k}{2}[1 - \ln(2)] \approx 0.15k$. The Akaike criterion, modified by substituting $f_k(\mathbf{Z}|\mathbf{Y})$ for $\widehat{f}_k(\mathbf{Z}|\widehat{\boldsymbol{\theta}}_k)$, chooses the hypothesis which maximizes $\ln(\widehat{L}_k) - 0.85k$.

If it is believed that the true model is a member of one of the approximating families under consideration, then an even better approach is feasible. For this situation, San Martini and Spezzaferri (1984) proposed a Bayesian prediction criterion for selecting the model with maximum posterior expected predictive accuracy. Reschenhofer (1996) implemented this approach for regression models. His simulation results suggested that substantial improvement in predictive accuracy over AIC is possible. Apparently, Forster (1997) is not opposed to this approach, but also is unwilling to specify a prior distribution for the unknown parameters. This seems to be an unnecessary limitation. Reschenhofer adopted a non-informative prior, yet still improved upon AIC.

Summing Up

We discussed how BTC and AIC work following an example. Bayesianism has been criticized for being subjective. One proclaimed strength of AIC is that it is not Bayesian. Unlike Bayesianism, it does not invoke prior probability. Despite the alleged strength of AIC as being non-Bayesian, we showed by contrast that AIC is, in fact, logically equivalent to BTC with a suitable choice of priors. In this connection, we discussed the relationship between Schwarz's Bayesian Criterion and BTC. We diagnosed the charge of subjectivity and argued that AIC approach has some defects as a research program.

7 Bibliography

1. Aitchison, J. (1975). "Goodness of Prediction Fit." *Biometrika*, **62**, 547–554.

2. Akaike, H. (1973). "Information Theory as an Extension of the Maximum Likelihood Principle," in B. N. Petrov and C. Csaki (Eds.), *Second International Symposium on Information Theory*, p.267–281, Budapest: Akademiai Kiado.

3. Akaike, H. (1978). "A Bayesian Analysis of the Minimum AIC Procedure." *Annals of the Institute of Statistical Mathematics*, **30**, 9–14.

4. Bandyopadhyay, P., R. Boik, and P. Basu (1996). "The Curve Fitting Problem: A Bayesian Approach." *Philosophy of Science*, **63** (supplement), 264–272.

5. Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

6. Forster, M. and E. Sober (1994). "How to Tell when Simple, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science* **45**, p.1–35.

7. Forster, M. (1997). "The New Science of Simplicity," in *Proceedings of the Tilburg Conference on Simplicity*, edited by Keuzenkamp, H et al, **9–11**, January, 1–28.

8. Jeffreys, H. (1961). *Theory of Probability* (3rd edition), London: Oxford University Press.

9. Linhart, H., & Zucchini, W. (1986). *Model Selection*, New York: John Wiley & Sons.

10. Reschenhofer, E. (1996). Prediction with Vague Knowledge. *Communications in Statistics — Theory and Methods*, **25**, 601–608.

11. San Martini, A., & Spezzaferri, F. (1984). A Predictive Model Selection Criterion. *Journal of the Royal Statistical Society*, Series B, **46**, 296–303.

12. Sawa, T. (1978). Information Criteria for Discriminating Among Alternative Regression Models. *Econometrica*, **46**, 1273–1291.

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

14. Weinberg, S. (1992). *Dreams of a Final Theory*, New York: Pantheon Books.