A RECONCILIATION OF THE INFORMATION AND POSTERIOR PROBABILITY CRITERIA FOR MODEL SELECTION

Gregory C. Chow*

Econometric Research Program
Research Memorandum No. 234

Revised, February, 1979

Abstract

This paper attempts to reconcile the information and the posterior probability criteria for selecting alternative models. It explains the logic of the two approaches, points out why they lead to different statistical procedures, and suggests which criterion should be used in practice.

* I would like to acknowledge financial support from the National Science Foundation.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey

A Reconciliation of the Information and Posterior Probability Criteria for Model Selection

Gregory C. Chow

1. INTRODUCTION

Two conflicting statistical criteria have been proposed for choosing among alternative models. One is the information criterion suggested by H. Akaike (1973, 1974). The second is the posterior probability criterion adopted by Bayesian statisticians including H. Jeffreys (1961), M. S. Geisel (1975), A. Zellner (1971), G. Schwarz (1978) and E. Leamer (1978), just to cite a few. These two criteria have led to different practical procedures to be used for selecting models which have different numbers of parameters. According to the information criterion estimated by Akaike (1973, 1974), given a set of sample observations, one should choose the model for which the maximum value of the log-likelihood function minus the number of its parameters is the highest. By the posterior probability criterion estimated by Schwarz (1978), one should choose the model for which the maximum value of the log-likelihood minus the number of its parameters times half the natural logarithm of the sample size is the highest. The purpose of this paper is to explain the logic of these two criteria, point out why they lead to different procedures, and suggest which criterion should be used in practice. A by-product of this paper is to clarify the role of a diffuse prior density function of the parameters in the selection of models by the posterior probability criterion.

In order to make clear the essential arguments of this paper, it is necessary to provide a brief restatement of each criterion in sections 2 and 3 before providing the final resolution in section 4.

2. THE INFORMATION CRITERION

Not to duplicate the expositions of the information criterion by Akaike (1973, 1974) and by T. Sawa (1978), I will merely state three key propositions to make clear the logic of this approach.

(P1) Assuming that $g(y)$ is the true probability density of a random vector of interest, and that a model $f(y|\theta)$ with a given parameter vector θ is selected to approximate $g(y)$, we will measure the goodness of the approximation by the information criterion

$$I[g;f(\cdot|\theta)] = E_y[\log g(y) - \log f(y|\theta)] = \int [\log g(y) - \log f(y|\theta)]g(y)dy \qquad (1)$$

If and only if $f(y|\theta) = g(y)$ almost everywhere (Kullback, 1959), the above criterion will be zero; otherwise it is positive. The information measure was adopted by S. Kullback and R. A. Leibler (1951), used by L. J. Savage (1954, pp. 50, 153, 235 ff) and studied at length by S. Kullback (1959).

(P2) Assuming that a sample of n observations $(y_1,\ldots,y_n) = Y$ is used to provide an estimate $\hat{\theta}(Y)$ of θ, we will measure the goodness of the estimated model $f(\cdot|\hat{\theta})$ by

$$E_{\hat{\theta}}\{I[g;f(\cdot|\hat{\theta})]\} \qquad (2)$$

where $\hat{\theta}$ is regarded as a constant when the expectation E_y is taken to form $I[g;f(\cdot|\hat{\theta})]$ as defined in (P1). Thus it is the mean of $I[g;f(\cdot|\hat{\theta})]$ over the sampling distribution of $\hat{\theta}$ which we will use to judge a particular model.

(P3) Assuming that the true model is $g(y) = f(y|\theta_o)$, that an approximate model $f(y|\theta)$ of k parameters is obtained by imposing restrictions on θ_o (such as specifying selected elements of θ_o to be zero), and that the estimate $\hat{\theta}$ is obtained by the method of maximum likelihood, H. Akaike has obtained an estimate of $-nE_{\hat{\theta}}\{I[g;f(\cdot|\hat{\theta})]\}$ as

$$-nE_{\hat{\theta}}\{I[g;f(\cdot|\hat{\theta})]\} \simeq -nE_y \log g(y) + \log L(Y,\hat{\theta}) - k \qquad (3)$$

where $L(Y,\theta)$ is the likelihood function for the approximate model $f(y|\theta)$.

Hence Akaike recommends choosing among several alternative models $f_1, f_2, \ldots, f_J$ the one which has the highest value of the maximum log-likelihood minus the number k of its parameters. Akaike (1974) has noted that his estimate (3) could be improved upon. For the selection of linear regression models, for example, Sawa (1978) has tried to improve upon the estimate (3).

To state clearly the situation appropriate for the application of the information criterion (estimated by Akaike's estimate (3) or another reasonable estimate), it is assumed that a researcher has specified several models $f_1(y|\theta_1), \ldots, f_J(y|\theta_J)$ to explain a random vector, that a sample of n observations $(y_1, \ldots, y_n) = Y$ has been obtained and that maximum likelihood estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ have been computed for the parameters of the models. The question is which estimated model will best fit the true density $g(y)$. This question is answered by using these models to predict a future observation y not yet available. One could measure the goodness of fit of each model by $E_y(y-\hat{y}_f)^2$ where $\hat{y}_f$ denotes the prediction of y obtained from the model $f(y|\hat{\theta})$. However, Akaike (1973) has advocated using a better measure according to the information criterion, i.e., $E_y[\log g(y) - \log f_j(y|\hat{\theta}_j)]$.

Although future observations are not yet available, we try to select a model $f_j(y|\hat{\theta}_j)$ which would do well, on the average, as judged by the information criterion should the future observations become available. The method of selection is provided by Akaike's estimate (3) or another reasonable estimate of $E_{\hat{\theta}}\{I[g;f(\cdot|\hat{\theta})]\}$.

3. THE POSTERIOR PROBABILITY CRITERION

To state the Jeffreys-Bayes posterior probability criterion, let $p(M_j)$ be the prior probability for model M_j to be correct, and $p(\theta|M_j)$ be the prior density for the k_j-dimensional parameter vector θ_j conditioned on M_j being correct. Assume that a random sample of n observations $(y_1,y_2\cdots y_n) = Y$ is available. By Bayes' theorem the posterior probability of the j-th model being correct is

$$p(M_j|Y) = \frac{p(M_j)p(Y|M_j)}{p(Y)} = \frac{p(M_j)p(Y|M_j)}{\sum_j p(M_j)p(Y|M_j)} \tag{4}$$

where

$$p(Y|M_j) = \int L_j(Y,\theta)p(\theta|M_j)d\theta \tag{5}$$

with $L_j(Y,\theta_j)$ denoting the likelihood function for the j-th model. Since $p(Y)$ in (4) is a common factor for all models, the model with the highest posterior probability of being correct is the one with the maximum

$$p(M_j)p(Y|M_j) = p(M_j)\int L_j(Y,\theta)p(\theta|M_j)d\theta \tag{6}$$

If the prior probabilities $p(M_j)$ are equal for the models, the one with the highest $p(Y|M_j)$ will be selected.

To demonstrate the application of this criterion, let us evaluate $p(Y|M_j)$ for large samples. Apply a well-known theorem of Jeffreys (1961, p. 193 ff), cited in Zellner (1977, pp. 31-33), on the posterior density $p(\theta|Y,M_j)$ of θ_j given model M_j:

$$p(\theta|Y,M_j) = \frac{L_j(Y,\theta)p(\theta|M_j)}{p(Y|M_j)} = (2\pi)^{-\frac{k_j}{2}}|S|^{\frac{1}{2}} e^{-\frac{1}{2}(\theta-\hat{\theta}_j)'S(\theta-\hat{\theta}_j)}[1+0(n^{-\frac{1}{2}})] \qquad (7)$$

where $\hat{\theta}_j$ is the maximum likelihood estimate of θ_j and the inverse covariance matrix $S = -(\partial^2 \log L_j/\partial\theta\partial\theta')_{\hat{\theta}} \equiv nR_j$ is of order n. (A rigorous proof of (7) is not our concern here since we are illustrating the logic of the posterior probability criterion and trying to explain why it gives a different answer.) Evaluating both sides of (7) at $\theta = \hat{\theta}_j$ and taking natural logarithms, we obtain

$$\log p(Y|M_j) = \log L_j(Y,\hat{\theta}_j) - \frac{k_j}{2}\log n - \frac{1}{2}\log|R_j| + \frac{k_j}{2}\log(2\pi) + \log p(\hat{\theta}_j|M_j) + 0(n^{-\frac{1}{2}}) \qquad (8)$$

If we retain only the two leading terms $\log L_j(Y,\hat{\theta}_j)$ and $-\frac{k_j}{2}\log n$ from (8), we obtain the formula of Schwarz (1978). By this formula, the model is selected if it has the highest value for the maximum log-likelihood minus the number of parameters times half of the logarithm of the sample size.

Two questions arise. First, why does formula (8) give a different procedure from formula (3) and which one should be used? Second, how should one choose a prior density $p_j(\theta|M_j)$ of the parameter vector for each model M_j if formula (8) is to be used? These questions will be answered in the next section.

4. A RESOLUTION OF THE CONFLICT

The reason why the posterior probability criterion, as exemplified by (8), gives a different statistical procedure from the information criterion as estimated by (3) is that it provides a solution to a different problem. In the last paragraph of section 2, we have stated the problem which the information criterion purports to solve. Inspection of formula (5) reveals that the posterior probability criterion will answer the following question. Given several models $f_1(y|\theta_1),\ldots,f_J(y|\theta_J)$, and given the corresponding prior density functions $p_j(\theta|M_j)$, which is the best model as judged by the evidence of the sample $(y_1,\ldots,y_n) = Y$? The answer is provided by the posterior probability of each model given the data Y , which is proportional to the likelihood $p(Y|M_j)$ of the data given the model under the assumption of equal prior probabilities $p(M_j)$. Note the difference between the definitions of the word "model" in the problems to be solved by the two criteria. For the posterior-probability criterion, the word "model" refers to $f_j(y|\theta_j)$ together with the prior information $p_j(\theta|M_j)$ on its parameter θ_j . For the information criterion, the "model" refers to $f_j(y|\hat{\theta}_j)$ where $\hat{\theta}_j$ is estimated by the sample data Y . Accordingly, the former uses the sample data Y to select a model which is specified by $f_j(y|\theta)$ and $p_j(\theta|M_j)$ prior to the sample, whereas the latter uses the sample data Y to estimate a model $f_j(y|\hat{\theta}_j)$ and asks which estimated model will best predict the future, yet unavailable observations. Thus using the posterior-probability criterion, one has specified a set of "pre-sample" models $[f_j(y|\theta);p_j(\theta|M_j)]$ and relies on the sample Y to select one among them. Using the information criterion, one has not merely specified a set of functions $f_j(y|\theta_j)$ but also estimated θ_j by $\hat{\theta}_j$ using the sample Y ;

one now asks which of these estimated models $f_j(y|\hat{\theta}_j)$ will predict best in the future.

It appears that the post-sample problem to be solved by the information criterion is more frequently the relevant one in practice. Since one has already observed the sample Y and bygones are bygone, why bother to ask the historical question as to which of the old, pre-sample models was the best? One probably cares more about the future. It is possible for a pre-sample model 1 to be better than a pre-sample model 2, as judged by the data Y, and for the post-sample model 1 to be worse for forecasting the future. This possibility can occur when model 1 is a model having a smaller number of parameters than model 2. With only a limited amount of pre-sample information or data, model 1 could have been more accurately estimated than model 2 and it did better in explaining the current sample Y. Once Y is available to estimate both models, the parameters of model 2 may now be sufficiently accurate to give better predictions in the future.

The distinction between a pre-sample model and a post-sample model has not been pointed out by the proponents of the posterior probability criterion. For example, Jeffreys (1961), Geisel (1975), Zellner (1971, Ch. 10) and Leamer (1978, Ch. 4) all implied that their method is designed to select a model for future prediction even when models having different numbers of parameters are involved. Schwarz (1978), in presenting his estimate of the posterior probability of a model being correct for large samples, stated that he was proposing an alternative formula to Akaike's for solving the same problem. Akaike (1978) asserted that he and Schwarz were trying to solve the same problem, and attempted to derive a formula close to his formula (3) by using the posterior probability criterion. This could be done, for example, by choosing the prior density $p(\hat{\theta}_j|M_j) = (2\pi)^{-\frac{k_j}{2}}|ne^{-2}R_j|^{\frac{1}{2}}$ in (8) to make the entire adjustment factor equal to $-k_j$ instead of $-\frac{k_j}{2}\log n$. The distinction between a pre-sample and a post-sample model becomes less impor-

tant, and thus the Bayesian method more useful, when the alternative models to be selected have the same number of parameters.

Bayesian statisticians including Jeffreys (1961), Pratt (1975) and Leamer (1978), among others, have recognized the difficult problem of choosing a prior distribution $p_j(\theta|M_j)$ for the parameters of each model to be used to compute $p(Y|M_j)$. The difficulty of this problem can be seen from equation (7), rewritten as

$$p(Y|M_j) = \frac{L_j(Y,\hat{\theta}_j)p(\hat{\theta}_j|M_j)}{p(\hat{\theta}_j|Y,M_j)}$$

$$\simeq L_j(Y,\hat{\theta}_j)p(\hat{\theta}_j|M_j) \cdot (2\pi)^{\frac{k_j}{2}}|nR_j|^{-\frac{1}{2}} \tag{9}$$

Observe that, given $L_j(Y,\hat{\theta}_j)$ and $p(\hat{\theta}_j|Y,M_j)$, $p(Y|M_j)$ is proportional to $p(\hat{\theta}_j|M_j)$. Thus one can change $p(Y|M_j)$ by a multiplicative factor simply by changing $p(\hat{\theta}_j|M_j)$ by that factor. If one wishes to use a diffuse prior density $p(\theta|M_j)$, many such densities are reasonable but they can give very different results. To illustrate, let $p(\theta|M_j)$ in (8) and (9) be k_j-variate normal with mean $\hat{\theta}_j$ (just for illustration) and covariance matrix $(\varepsilon R_j)^{-1}$. Equation (8) will become

$$\log p(Y|M_j) = \log L_j(Y,\hat{\theta}_j) - \frac{1}{2} k_j \log\left(\frac{n}{\varepsilon}\right) + O(n^{-\frac{1}{2}}) \tag{10}$$

The adjustment factor suggested by the formula of Schwarz (1978) will be changed from $-\frac{1}{2} k_j \log n$ to $-\frac{1}{2} k_j \log(\frac{n}{\varepsilon})$. There is no reason why ε might not be 1, 2.3 or 4.1, making the formula useless in practice.[1]

The reason for the difficulty in choosing a robust prior density function $p(\theta|M_j)$ for the model selection problem is that the "model" to be judged by the sample data Y is precisely defined by this prior density together with the function $f_j(y|\theta)$. Varying the function $p(\theta|M_j)$ will

vary significantly the "model" to be judged. Therefore, it does not make sense to look for a diffuse prior in this situation. One might be tempted to resolve this dilemma by using a part Y_1 of the sample $Y = (Y_1\ Y_2)$ to obtain a preliminary $p(\theta|Y_1,M_j)$ from a diffuse $p(\theta|M_j)$, and then using the remaining data Y_2 to judge the "model" now specified by $p(\theta|Y_1,M_j)$ together with the function $f_j(y|\theta)$. This suggestion can certainly be carried out, but it will answer the question whether the second "model" was good as judged by the data Y_2, and not whether the original model with a diffuse prior was good. Nor will it answer the more interesting question whether the model estimated by using all the data Y will be good.

In conclusion, we have observed that the information and posterior probability criteria, while both correct, are appropriate for answering different questions because the "model" to be selected in each case is different. In general, the concept of a "model" in a model selection problem cannot be usefully defined by the mathematical function alone, without specifying the amount of information available. Both criteria have included an amount of information in specifying the "models" to be selected, be it the sample data Y or the prior density $p_j(\theta|M_j)$. Consequently, it seems uninteresting to ask which of the functions $f_j(y|\theta_j)$ is the correct one when θ_j include different numbers of parameters. For example, in the problem of selection from two linear regression models, one with X_1 alone and the other with both X_1 and X_2 as explanatory variables, the latter must be closer to the true function for practical applications in the social sciences. The former model includes the unreasonable assumption that the coefficients β_2 of X_2 are <u>exactly</u> zero, or X_2 have absolutely no effect. The interesting question is not whether $X_1\beta_1 + X_2\beta_2$ is closer to the true regression function than $X_1\beta_1$, as we know it is; it is rather which function, together with the limited data Y at our disposal, can be used to construct a better model for predicting the future.

Footnote

[1]Leamer (1978, pp. 111-112), in attempting to find a diffuse prior $p_j(\theta|M_j)$ to evaluate $p(Y|M_j)$ for a linear regression model, obtained three unreasonable answers (p. 111). These unreasonable answers are due to the failure to specify a prior density $p_j(\theta|M_j)$ which is consistent with the units of measurement for the explanatory variables X_j. If we double all the explanatory variables by changing their units, $X_j'X_j$ will become $4X_j'X_j$; the inverse covariance matrix N^*_j of the prior density for θ_j should also become $4N^*_j$. To maintain the consistency in units, one should let $N^*_j=\varepsilon\cdot\frac{1}{n}X_j'X_j$, which corresponds to the specification εR_j as the inverse covariance matrix of $p(\theta|M_j)$ in our equations (8) and (9). This would rule out the unreasonable answers of Leamer (p.111). However, as we have just pointed out, an indeterminacy in specifying a value for ε still remains. This indeterminacy corresponds to the indeterminacy of the constant c in equation (4.16) of Leamer (p. 112). By arbitrarily fixing c, one can make the answer come out anyway one wishes, making Leamer's equation (4.16) useless in practice.

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," Proceedings of the 2nd International Symposium on Information Theory, ed. by B. N. Petrov and F. Csáki, Budapest: Akademiai Kiadó, 267-281.

_____ (1974), "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, AC-19, 716-723.

_____ (1978), "A Bayesian Analysis of the Minimum AIC Procedure," Annals of the Institute of Statistical Mathematics, Part A, 30, 9-14.

Geisel, M. S. (1975), "Bayesian Comparisons of Simple Macroeconomic Models," in Studies in Bayesian Econometrics and Statistics, ed. by S. E. Feinberg and A. Zellner. Amsterdam: North-Holland, 227-256.

Jeffreys, H. (1961), Theory of Probability, (3rd ed.) Oxford: Clarendon.

Kullback, S. (1959), Information Theory and Statistics, New York: John Wiley and Sons.

Kullback, S. and R. A. Leibler (1951), "On Information and Sufficiency," Annals of Mathematical Statistics, 22, 79-86.

Leamer, E. (1978), Specification Searches, New York: John Wiley and Sons.

Pratt, J. W. (1975), "Comments," in Studies in Bayesian Econometrics and Statistics, ed. by S. E. Feinberg and A. Zellner. Amsterdam: North-Holland, 71-73.

Savage, L. J. (1954), The Foundations of Statistics, New York: John Wiley and Sons.

Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models," Econometrica, 46, 1273-1292.

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

Zellner, A. (1971), An Introduction to Bayesian Influence in Econometrics, New York: John Wiley and Sons.