Reprinted from *Bayesian Statistics*, J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds., University Press, Valencia, Spain, 1980, 1-13 by permission from J. M. Bernardo

Likelihood and the Bayes procedure

HIROTUGU AKAIKE

The Institute of Statistical Mathematics, Tokyo

SUMMARY

In this paper the likelihood function is considered to be the primary source of the objectivity of a Bayesian method. The necessity of using the expected behavior of the likelihood function for the choice of the prior distribution is emphasized. Numerical examples, including seasonal adjustment of time series, are given to illustrate the practical utility of the common-sense approach to Bayesian statistics proposed in this paper.

Keywords: LIKELIHOOD; BAYES PROCEDURE; AIC; SEASONAL ADJUSTMENT.

1. INTRODUCTION

The view that the Bayesian approach to statistical inference is useful, practically as well as conceptually, is now widely accepted. Nevertheless we must also accept the fact that there still remain some conceptual confusions about the Bayes procedure. Although many strong impetuses for the use of the procedure came from the subjective theory of probability, it seems that the confusions are also caused by the subjective interpretation of the procedure.

By looking through the works on the Bayes procedure by subjectivists, it quickly becomes clear that there is not much discussion of the concept of likelihood. The subjective theory of probability is used only to justify the use of the prior distribution of the parameters of a data distribution. It is almost trivial to see that no practically useful Bayes procedure is defined without the use of the likelihood function, while the likelihood function can be defined without the prior distribution. Thus the data distribution represents the basic part of our prior information and the Bayes procedure gives only one specific way of utilizing the information supplied by data through the likelihood function.

From this point of view there is nothing special about the choice of prior distributions to differentiate it from the design of ordinary statistical procedu-

res such as the choice of the sampling procedure in a sample survey and the choice of the spectrum window in the spectrum analysis of a time series.

In this paper we first discuss some conceptual confusions with the Bayes procedure which we believe to be due to the subjective interpretation of the procedure. We argue that it is necessary to recognize the limitation of the subjective theory and put more emphasis on the concept of likelihood. We take the position of regarding the Bayes procedure as one possible way of utilizing the information provided by the likelihood function. Once such an attitude towards the Bayes procedure is accepted we can freely develop Bayesian models simply by representing a particular preference of the parameters by a prior distribution. The goodness of the prior distribution can then be checked by evaluating expected performances of the corresponding Bayes procedure in various conceivable situations.

We demonstrate the use of this type of approach by developing a general Bayesian model for the analysis of linear relations between variables. The model contains as special cases the basic models of those estimation procedures such as the Stein estimator, ridge regression, Shiller's distributes lag estimator and O'Hagan's localized regression. Numerical examples are given to illustrate the practical utility of some quasi-Bayesian procedures developed for these models and for a more conventional model of polynomial regression. The result of application to the seasonal adjustment of time series seems particularly interesting as the model contains twice as many parameters as the number of the observations.

2. CONCEPTUAL DIFFICULTIES OF THE SUBJECTIVE APPROACH

Significant impetus for the advancement of Bayesian statistics has come from the side of the subjective theory of probability. This is natural as every statistical procedure may be viewed as a formulation of the psychological process of information processing and evaluation by a skilful researcher. In spite of the significant contribution of the subjective theory of probability to clarifying the nature of the psychological aspect of this process, several conceptual difficulties remain with the theory. Here we discuss some difficulties, which we believe to be misconceptions, related to the Bayes procedure and clear the way for the development of practically useful Bayesian methods.

2.1. *Rationality and Savage's axiom*

It is sometimes said that a rational person must behave as if he has a clearly defined system of subjective probabilities of uncertain events. This is often ascribed to Savage (1954) who developed a theory of personal probability by axiomatizing the preference behavior of a person under uncertainty. Un-

fortunately the very first postulate P1 of Savage, which assumes the linear ordering of the preference, excludes the real difficulty of preference. This can be explained by the following simple example.

Consider a young boy who wants to choose a girl as his wife. His preference is based on the three characteristics, H, I and L. Here H stands for health, I for intelligence and L for looks. Each characteristic is ranked by the numbers 1, 2, and 3, with higher number denoting higher rank. The difference of ranks by 1 is marginal and the difference by 2 means a significant difference. Denote by $R_i = (H_i, I_i, L_i)$ the vector of the ranks of the characteristics of the i^{th} girl. Being uncertain about the relative importance of these characteristics in his future life, he ignores the marginal differences and pays attention only to the significant differences. Thus his preference is defined by the following scheme:

$$R_i \leq R_j, \textit{ i.e.}, \text{ the } j^{th} \text{ girl is preferred to the } i^{th} \text{ girl,}$$

$$\text{iff } C_i \leq C_j \text{ for the characteristic } C$$

$$\text{for which } |C_j - C_i| \text{ is maximum.}$$

Now he has three girl friends ($i = 1, 2, 3$) whose R_i's are defined by $R_1 = (1, 2, 3)$, $R_2 = (3, 1, 2)$ and $R_3 = (2, 3, 1)$, respectively. Obviously it holds that

$$R_1 \leq R_2, R_2 \leq R_3 \text{ and } R_3 \leq R_1,$$

which shows that his natural preference scheme does not satisfy the postulate P1 of Savage.

It is the difficulty of this type of preference that make us feel the need of a horoscope or some other help in making the decision in a real life situation. Since Savage's system excludes the possibility of this type of difficulty, the corresponding theory of personal probability cannot tell how we should treat the difficulty. The exact characterization of Savage's theory is then a theory of one particular aspect of preference and there is no compelling reason to demand that a rational person's preference should be represented by a single system of subjective probability. Wolfowitz (1962) presents a pertinent discussion of this point. Thus to justify the use of a system of personal probability one must prove its adequacy by some means. Certainly the proof cannot be found within the particular system of personal probability itself.

2.2. *The role of parameters in a Bayesian modeling*

The subjective theory of probability of De Finetti demands that the probability distribution or the expectations of the uncertain events of interest

should completely be specified (de Finetti, 1974b, p. 87). If we accept this demand and decide to use the Bayes procedure, all we have to do is to compute $p(y|x)$, the probability of an event y conditional on a given set of data x. The theory only asserts that the necessary probability distribution should be there and does not consider the special role played by the parameters in constructing a statistical model or the probability distribution. De Finetti (1974a, p. 125) even rejects the concept of a parameter as metaphysical, unless it is a decidable event.

That the concept of parameter cannot be eliminated is shown by the simple example of the binomial experiment where the probability of occurence of a head in a coin tossing is considered. The concept of independent trials with a fixed probability of head is unacceptable by the subjective theory of probability of de Finetti and the solution is sought in the concept of exchangeability (de Finetti, 1975, pp. 211-218). The difficulty is caused by the fact that the probability of a head, which must be decided, plays the role of a parameter that is not actually decidable (Akaike, 1979b).

We may use the theory of probability to develop some understanding of what we psychologically expect of the parameters of a statistical model. Consider a random variable x and the observations x_1, x_2, ...of some related events. We expect that a parameter θ exhausts the information about x to be gained through the observations x_1, x_2... The probabilistic expression of this expectation is given by

$$p(x|\theta, x_1, x_2, \ldots) = p(x|\theta), \tag{2.1}$$

where $p(x|z_1, z_2, \ldots)$ denotes the distribution of x conditional on $z_1, z_2, \ldots$. To allow this type of discussion we must consider θ as a random variable as is advocated by Kudo (1973). The formula (2.1) then gives a very natural characterization of the parameters as a condensed representation of the information contained in the observations, i.e., once θ is known no further observations can improve our predictions on x. Thus we want to know the value of θ. Actually de Finetti's discussion of the exchangeable distribution of the binomial experiment has given a proof of the existence of such a variable.

Although the above characterization of a parameter is interesting, in the statistical model building for inference the order of reasoning is reserved. The prior information first suggests what type of parameterization of the data distribution $p(x|\theta)$ should be used. The prior distribution $\pi(\theta)$, if at all specified, represents only a part of the prior information. To take the parameters as something prespecified and assume that the prior distribution can or should be determined independently of the data distribution constitutes a serious misconception about the inferencial use of the Bayes procedure.

2.3. *Likelihood principle and the Bayes procedure*

It has often been claimed that the likelihood principle, which demands that the statistical inference should be identical if the likelihood function is identical, is a direct consequence of the Bayesian approach; see, for example, Savage (1962, p. 17). In the example of coin tossing, if we denote the probability of head by θ and assume the independence and homogeneity of the tossings, we have

$$p(x|\theta) = C\,\theta^x (1-\theta)^{n-x}$$

as the likelihood of θ when x heads appeared in n tossings. It is argued that there is no difference in the inference through the Bayes procedure if the above likelihood is obtained as the result of n tosses, with n predetermined, or as the result of tossing continued until x heads appeared, with x predetermined.

This seemingly innocuous argument is against the principle of rationality of the subjective theory of probability which suggests that the choice of a statistical decision be based on its expected utility. The expected behavior of the likelihood function $p(x|\theta)$ is certainly different for the two schemes of the coin tossing and it is irrational to adopt one and the same prior distribution $\pi(\theta)$, irrespectively of the expected difference of the statistical behavior of the likelihood functions.

To clarify the nature of the confusion by a concrete example, consider the use of the posterior distribution

$$C\,\theta^x(1-\theta)^{n-x}\,\pi(\theta)$$

as an estimate of the probability distribution of the result y of the next toss, where $y = 1$ for head and 0 otherwise. The predictive distributions are defined as the averages of the data distribution $p(y|\theta)$ with respect to the posterior distributions of θ. These will be denoted by $p(y|x)$ and $p(y|n)$ to indicate that x and n are the realizations of the random variables, respectively. They are defined by

$$p(y|*) = C\int_0^1 \theta^{x+y}(1-\theta)^{n+1-x-y}\pi(\theta)\mathrm{d}\theta,$$

where $*$ stands for either x or n. When the "true" value of θ is θ_0 the goodness of $p(y|*)$ as an estimate of the true distribution $p(y|\theta_0) = \theta_0^y (1-\theta_0)^{1-y}$ can be measured by the entropy of $p(y|\theta_0)$ with respect to $p(y|x)$ or $p(y|n)$ which is defined by

$$\mathrm{B}\{p(\cdot|\theta_o), p(\cdot|*)\}$$
$$= -\Sigma_{y=0}^{1}\left\{\frac{p(y|\theta_o)}{p(y|*)}\right\}\log\left\{\frac{p(y|\theta_o)}{p(y|*)}\right\}p(y|*)$$

The larger the entropy the better is the approximation of $p(\cdot|*)$ to $p(\cdot|\theta_o)$. Before we observe x or n we evaluate $\mathrm{E}_* \mathrm{B}\{p(\cdot|\theta_o), p(\cdot|*)\}$ for some possible values of θ_o, where E_* denotes the expectation with respect to the distribution of $*$ defined with $\theta = \theta_o$. We have

$$\mathrm{E}_x \mathrm{B}\{p(\cdot|\theta_o), p(\cdot|x)\}$$
$$= \Sigma_{y=0}^{1} p(y|\theta_o)\,\Sigma_{x=0}^{n} \log\left\{\frac{p(y|x)}{p(y|\theta_o)}\right\}{}_nC_x\,\theta_o^x(1-\theta_o)^{n-x}$$

and

$$\mathrm{E}_n \mathrm{B}\{p(\cdot|\theta_o), p(\cdot|n)\}$$
$$= \Sigma_{y=0}^{1} p(y|\theta_o)\,\Sigma_{n=x}^{\infty} \log\left\{\frac{p(y|n)}{p(y|\theta_o)}\right\}{}_{n-1}C_{x-1}\,\theta_o^x(1-\theta_o)^{n-x}.$$

Obviously we have no reason to expect that these two quantities will take one and the same value and, at least for that matter, there is no reason for us to assume one and the same prior distribution $\pi(\theta)$ for both cases.

3. LIKELIHOOD AS THE SOURCE OF OBJECTIVITY

The discussion in the preceding section illustrates both the subjective and objective elements in the Bayesian approach to statistical inference. It is subjective because a statistical inference procedure is designed to satisfy a subjectively chosen objective. The choice of the data distribution is particularly subjective and the prior distribution reflects the object of the inference which is often expressed in the form of a psychological expectation.

What is then objective with the procedure ? The objectivity stems from the dependence on the data which is a production of the outside world. This objectivity is fed into the Bayes procedure through the likelihood function. Since $\mathrm{B}\{p_o(\cdot), p(\cdot|\theta)\} = \mathrm{E}_x \log p(x|\theta) - \mathrm{E}_x \log p_o(x)$, we can see that, ignoring the additive constant $\mathrm{E}_x \log p_o(x)$, the log likelihood $\log p(x|\theta)$ is a natural estimate of the entropy of $p_o(\cdot)$ with respect to $p(\cdot|\theta)$. Here E_x denotes the expectation with respect to the distribution $p_o(\cdot)$ of x. Thus the likelihood $p(x|\theta)$ represents an objective measure of the goodness, as measured by x, of

$p(\cdot|\theta)$ as an approximation to $p_o(\cdot)$. This fact forms the basis of the practical utility of the Bayes procedure even for the family $\{p(\cdot|\theta)\}$ which is chosen subjectively and does not contain the true distribution of x.

The likelihood function $p(x|\theta)$ is the basic device for the extraction or condensation of the information supplied by the data x. The role of the prior distribution $\pi(\theta)$ is to aid further condensation of the information supplied by the likelihood function $p(x|\theta)$ through the introduction of some particular preference of the parameters. By evaluating the expected entropy of the true distribution with respect to the predictive distribution specified by a posterior distribution we can extend the concepts of bias and variance to the posterior distribution (Akaike, 1978a). If we try to keep a balance between the bias and variance, we cannot ignore the influence of the statistical behavior of the likelihood function on the choice of our prior distribution. Some of the conflicts between the conventional and Bayesian statistics are caused by ignoring the possible dependence of the choice of the prior distribution, or even the choice of the basic data distribution, on the number of available observations which influences the behaviour of the likelihood function; see, for example, Lindley (1957), Schwarz (1978) and Akaike (1978b).

4. A GENERAL BAYESIAN MODELING FOR LINEAR PROBLEMS

In this section we demonstrate the practical utility of the point of view discussed in the preceding section through the discussion of a general Bayesian model for the analysis of linear problems. The basic idea here may be characterized as the common-sense approach to Bayesian statistics.

Consider the analysis of the linear relation between the vector of observations $y = [y(1),\ldots,y(N)]'$ and the vectors of the independent variables $x_i = [x_i(1), x_i(2),\ldots x_i(N)]'$ $(i = 1,2,..,K)$, where $'$ denotes transposition. The method of least squares leads to the minimization of

$$L(a) = \Sigma_{j=1}^{N} [y(j) - \Sigma_{i=1}^{K} a_i x_i(j)]^2. \tag{4.1}$$

We know, when K is large compared with N or when the matrix $X = [x_1, x_2,\ldots,x_K]$ is ill-conditioned the least squares estimates behave badly. To control this we introduce some preference on the values of the parameters and try to minimize

$$L(a) + \mu \|a - a_0\|_R^2 \tag{4.2}$$

where a_0 denotes a particular vector of parameters $[a_{01}, a_{02},\ldots,a_{0K}]'$, $\| \|_R^2$ the norm defined by a positive definite matrix R, and μ a positive constant. The use of this type of constrained least squares for the solution of

an ill-posed problem is wellknown; see, for example, Tihonov (1965).

The difficulty with the application of this method of constrained least squares is in the choice of the value of μ. To solve this we transform the problem into the maximization of

$$\ell(a) = \exp\left\{-(1/2\sigma^2)\,[L(a) + \mu\|a - a_0\|_R^2]\right\},$$

where temporarily σ^2 is assumed to be known. Since we have

$$\ell(a) = \exp\left[-(1/2\sigma^2)\,L(a)\right] \exp\left[-(\mu/2\sigma^2)\,\|a-a_0\|_R^2\right],$$

we can see that the solution of the constrained least squares problem is now given as the mean of the posterior distribution defined by the data distribution

$$f(y|\sigma^2,a) = (1/2\pi)^{N/2}(1/\sigma)^N \exp\left[-(1/2\sigma^2)\,L(a)\right], \tag{4.3}$$

and the prior distribution

$$\pi(a|d) = (1/2\pi)^{K/2}\,(1/\sigma)^K \exp\left[-(d^2/2\sigma^2)\,\|a-a_0\|_R^2\right], \tag{4.4}$$

where $d^2 = \mu$. By properly choosing X, a_0 and R, we can get many practically useful models. Particularly, we will restrict our attention to the case where $\|a - a_0\|_R^2$ is defined by

$$\|a - a\|_R^2 = \|c_0 - Da\|^2, \tag{4.5}$$

where D is a properly chosen matrix, $c_0 = Da_0$ and $\|v\|^2$ denotes the sum of squares of the components of v. In this case the posterior mean of the vector parameter a is obtained by minimizing $\|z(a|d)\|^2$ of the vector $z(a|d)$ defined by

$$z(a|d) = \begin{bmatrix} y(1) \\ y(2) \\ \cdot \\ \cdot \\ \cdot \\ y(\mathrm{N}) \\ dc_0(1) \\ dc_0(2) \\ \cdot \\ \cdot \\ dc_0(L) \end{bmatrix} - \begin{bmatrix} \\ \\ \\ X \\ \\ \\ \\ aD \\ \\ \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ \\ \\ a(\mathrm{K}) \end{bmatrix} \tag{4.6}$$

Examples.

a. *Stein type shrunken estimator*

This is defined by putting $L = N$, $D = X$ and $c_o = 0$, the zero vector. The case with $K = N$ and $X = I_{NxN}$ corresponds to the original problem of estimation of the mean vector of a multivariate Gaussian distribution treated by Stein. By putting c_o equal to the vector of the parameters obtained from some similar former observations, we can realize a reasonable use of the prior information.

b. *Ridge regression*

This is defined by putting $L = K$, $D = I_{KxK}$ and $c_o = 0$.

c. *Shiller's distributed lag estimator*

Shiller (1973) developed a procedure for the estimation of a smoothly changing impulse response sequence. In this case $[y(1), y(2), \ldots, y(N)]$ is obtained as the time series of the output of a constant linear system under the input $u(j)$. X is defined by $x_i(j) = u(j-i+1)$ and $c_o = 0$.

D is put equal to

$$D_1 = \begin{bmatrix} \alpha & & & & & & \\ -1 & 1 & & & & & \\ & -1 & 1 & & 0 & & \\ & & \cdot & \cdot & & & \\ & 0 & & \cdot & \cdot & & \\ & & & & \cdot & \cdot & \\ & & & & & -1 & 1 \end{bmatrix}$$

or

$$D_2 = \begin{bmatrix} \alpha & & & & & & & \\ -\beta & \beta & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & 1 & & 0 & & \\ & & \cdot & \cdot & \cdot & & & \\ & & & \cdot & \cdot & \cdot & & \\ & 0 & & & \cdot & \cdot & \cdot & \\ & & & & & 1 & -2 & 1 \end{bmatrix}$$

where α and β are properly chosen constants. D_1 controls the first order differences of $a(j)$ and D_2 the second order differences.

d. *Localized regression of O'Hagan.*

O'Hagan (1978) introduced an interesting Bayesian model for the estimation of the locally gradually changing regression of a time series $y(i)$ on $x(i)$. Our model corresponding to O'Hagan's is given by putting $K = N$, $c_o = 0$ and

$$X = \begin{bmatrix} x(1) & & & & \\ & x(2) & & 0 & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & 0 & & & x(N) \end{bmatrix}.$$

D is put equal to D_1 or D_2 of the above example or

$$D_3 = \begin{bmatrix} \alpha & & & & & & & \\ -\beta & \beta & & & & & & \\ \gamma & -2\gamma & \gamma & & & & 0 & \\ -1 & 3 & -3 & 1 & & & & \\ & -1 & 3 & -3 & 1 & & & \\ & & \cdot & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \cdot & \\ & 0 & & & \cdot & \cdot & \cdot & \cdot \\ & & & & & -1 & 3 & -3 & 1 \end{bmatrix}$$

One particularly interesting model is obtained by putting $x(i) = 1$ $(i = 1,2, \ldots, N)$. The number of parameters in this model is equal to the number of observations $y(i)$.

e. *Locally smooth trend fitting*

For a time series $y(i)$, by putting $c_o = 0$ and $D = D_k X$ where D_k is as given in the preceding examples, we get a model for the fitting of a smooth trend curve. One special choice of X is given by $X = I_{N \times N}$. We will call the model defined with $X = I_{N \times N}$ and $D = D_k$ the model of locally smooth trend of k^{th} order.

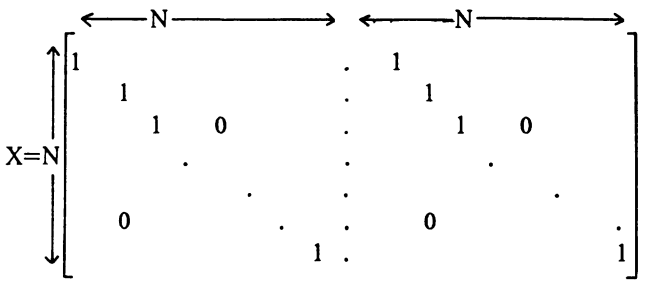
f. *Bayesian seasonal adjustment*

We consider the decomposition of the monthly observations $y(i)$ for M years, where $i = 12m + j$ $(j = 1,2, \ldots, 12, m = 0,1, \ldots, M\text{-}1)$, into the form

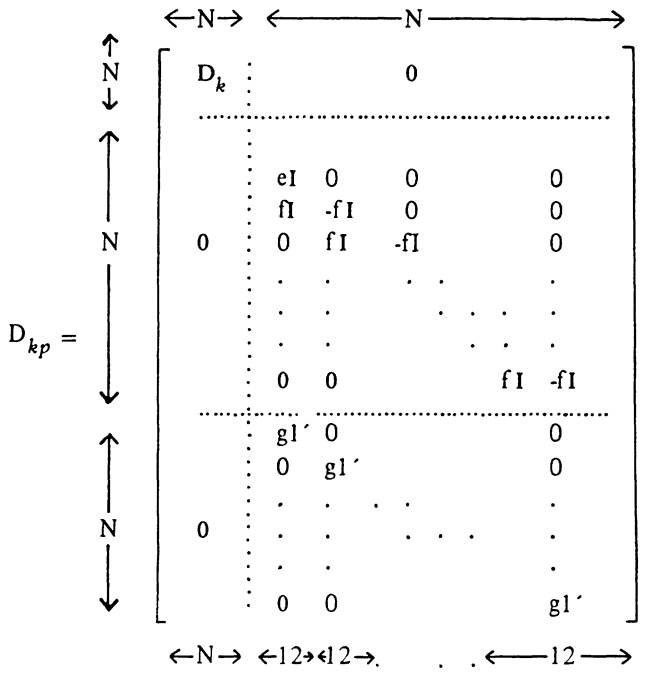
$$y(i) = T_i + S_i + I_i,$$

where T_i denotes the trend, S_i the seasonal and I_i the irregular component. For this problem we put $K = 2N$ $(N = 12M)$ and define $a = (T_1, T_2, \ldots, T_N, S_1, S_2, \ldots, S_N)$ and put $c_o = 0$.

The matrix X is defined by



and D by



where D_k is one of those defined in the preceding examples, $I = I_{12x12}$, $1' = (1, 1, \ldots, 1)$, and e, f, g are properly chosen constants.

A notable characteristic of this model is that it has twice as many parameters as the number of observations. This constitutes a typical many parameters problem which cannot be handled by the ordinary unconstrained least squares or the method of maximum likelihood.

The fundamental problem in applying these models to real data is the choice of the constant d. Assuming that other constants are specified, the decision on d is equivalent to the decision on the prior distribution of a. From (4.2) the choice of d, or μ, determines the relative weight of the additional term $\| a - a_0 \|_R^2$ against $L(a)$, the sum of squares of the residuals. When a_0 is not exactly equal to the true value of a, we expect that the bias of the estimate increases as d is increased but the variance decreases. It is natural to try to keep a balance between these two factors. To realize this it is necessary not to specify d uniquely but use the information supplied by the likelihood function or $L(a)$.

In the Bayesian terminology this is to consider d as a hyperparameter which has its own prior distribution. Now it is obvious that by considering d as a hyperparameter we are trying to use the information supplied by the likelihood function for the determination of d. This observation suggests that

a proper choice of the prior distribution to be used in an inferential situation can only be realized through the analysis of the statistical characteristics of the related likelihood function. The infinite digression of considering the priors of priors can only be stopped by the analysis of the expected output at each stage, which is determined by the behavior of the likelihood function.

Incidentally, the present observation shows why the conventional subjectivist doctrine of assuming the determination of the prior distribution of the parameters independently of the related likelihood function was not strictly followed by the research workers dealing with real inference problems. This point is discussed as the Bayes / Non-Bayes compromise by Good (1965). We take here the very flexible attitude towards the Bayes procedure to consider it only as one possibility of utilizing the information supplied by the likelihood function. Thus we consider that any practically useful statistical procedure which utilizes the information supplied by the likelihood function should not be rejected only because it is non-Bayesian. It is not the dogmatic exclusion of other procedures but the explicit proposal of useful models that proves the advantage of the Bayesian approach over the conventional statistics.

5. NUMERICAL EXAMPLES

To show that our discussion in the preceding sections is not vacuous, here we show some numerical examples. These were obtained by Bayesian modelings but with the help of some procedures which are not strictly Bayesian. The first three examples are concerned with the models discussed in the preceding section. The last one is an example of polynomial fitting and is included to show the feasibility of a Bayesian modeling with the aid of an information criterion (AIC) to deal with the difficulty of choosing a prior distribution for a multimodel situation where the models are with different number of parameters.

For the first three examples the essential statistic used for the determination of the parameter d in (4.4) is the likelihood of the model specified by the prior distribution. We consider the marginal likelihood of (d,σ^2) defined by

$$L(d,\sigma^2) = \int f(y|\sigma^2,a)\,\pi(a|d)\,da,$$

where $f(y|\sigma^2,a)$ and $\pi(a|d)$ are given by (4.3) and (4.4), respectively. If we assume (4.5) and put $c_0 = 0$ we get

$$L(d,\sigma^2) = (1/2\pi)^{N/2}(1/\sigma)^N \exp\left[-(1/2\sigma^2)\| z(a_*|d)\|^2\right]$$

$$\cdot \| d^2D'D \|^{1/2} \| d^2D'D + X'X \|^{-1/2},$$

where $\| z(a_\bullet|d) \|^2$ denotes the minimun of $\| z(a|d) \|^2$ with $z(a|d)$ defined by (4.6). Instead of developing a prior distribution of (d,σ^2) we consider the use of the procedure which chooses a model with the maximum marginal likelihood. This is called the method of type II maximum likelihood by Good (1965). For a given d, the maximum with respect to σ^2 is attained at

$$\sigma_d^2 = (1/N) \| z(a_\bullet|d) \|^2.$$

For the case of practical applications, we consider a finite set of possible values $(d_1, d_2, \ldots, d_l)$ of d and choose the one that maximizes $L(d,\sigma_d^2)$. Since we are familiar with the use of minus twice the log likelihood, we propose to minimize

$$\text{ABIC} = (-2) \log L(d, \sigma_d^2)$$

$$= N \log [\, 1/N \| z(a_\bullet|d) \|^2 \,] + \log \| d^2 D'D + X'X \|$$

$$- \log \| d^2 D'D \| + \text{const},$$

where ABIC stands for "a Bayesian information criterion". When different D's are not considered, the term $\log \| d^2 D'D \|$ may be replaced by $2K \log d$, where K is the dimension of the vector a.

In the last example we demonstrate the practical utility of $\exp(-\frac{1}{2}\text{AIC})$ as the definition of the likelihood of a model specified by the maximum likelihood estimates of the parameters. Here AIC is by definition (Akaike, 1974)

$$\text{AIC} = (-2) \log (\text{maximum likelihood}) + 2 (\text{number of free parameters}).$$

This definition allows a very practical procedure of developing a Bayesian type approach to the situation where several models with different numbers of parameters are considered.

The general definition of ABIC of a model with hyperparameters determined by the method of type II maximum likelihood would have been ABIC = (-2) log (maximum marginal likelihood) + 2 (number of adjusted hyperparameters). In the examples treated in this paper the numbers of the adjusted hyperparameters are identical within the models being compared and their influence on the maximum marginal likelihoods is ignored.

Examples

a. *Distributed lag estimation*

We did a simulation with the second example of Shiller (1973, p. 783). The result is illustrated in Table 1. This result was obtained by using the model

c of the precedin section with $N = 40$, $K = 20$ and $D = D_2$ with $\alpha = \beta = 0$. Considering that this is a limiting situation with non-zero α and β, ABIC was defined by

$$\text{ABIC} = N \log [\, (1/N) \mid z(a_* \mid d) \mid^2]$$

$$+ \log \mid d^2 D'D + X'X \mid - 2K \log d,$$

and the ABIC was minimized over $d = 5.0, 2.5, 1.25, 0.625, 0.3125$. the values of the ABIC at these d's were -43.4, -51.5, -52.7, -45.0, -30.9, respectively. The minimum, -52.7, was attained at $d = 1.25$ and corresponding estimates of the parameters are given in Table 1 along with the theoretical values and the least squares estimates. By taking a properly weighted average of the results with different d's we may get a procedure which has smaller sampling variability, but it seems that the present simple procedure is almost sufficient for many practical applications.

TABLA 1
Example of distributed lag estimation

i	1	2	3	4	5
Theoretical	.000	.000	.001	.004	.018
Bayes	-.009	-.003	.004	.009	.017
Least squares	-.010	.021	-.045	.037	.078
i	6	7	8	9	10
Theoretical	.054	.130	.242	.352	.399
Bayes	-.051	.134	.242	.345	.395
Least squares	-.074	.255	.113	.462	.334
i	11	12	13	14	15
Theoretical	.352	.242	.130	.054	.018
Bayes	.362	.257	.134	.052	.012
Least squares	.359	.329	.046	.072	.042
i	16	17	18	19	20
Theoretical	.004	.001	.000	.000	.000
Bayes	-.001	-.015	.006	.035	-.018
Least squares	-.018	-.050	.065	-.008	-.008

b. *Locally smooth trend fitting*

In this example the original data $y(i)$ ($i = 1, 2, \ldots, 30$) were generated by

the relation

$$y(i) = 4 \exp [- (1/2) ((i-5)/4)^2] + z(i),$$

where z(i)'s are independently and identically distributed as $N(0,1)$. Twelve models of locally smooth trend of k^{th} order defined by the model e of the preceding section with $d = 2^{8-j}$ ($j = 1, 2, \ldots, 12$) were tried with $k = 1, 2, 3$. The constants α, β, and γ of the D_k's were all put equal to 0.001. The ABIC was defined by

$$\text{ABIC} = N \log [(1/N) \| z(a_* | d) \|^2] + \log \| d^2 D'D + X'X \|$$

$$-\log \| d^2 D'D \|.$$

The minimum of ABIC was attained at $k = 1$ and $d = 2.0$. The original data, the theoretical trend and some of the estimated trends are illustrated in Fig. 1. In this figure SSDEV stands for the sum of squares of deviations of the estimates from the theoretical. It can be seen that the present procedure can produce meaningful results even with these rather noisy observations. In the figures *ID* stands for k.

c. *Seasonal adjustment*

In this case the model f was applied to various artificial and real time series of length six years, i.e., $N = 72$. The constants of D_k in D_{kp} were the same as in the preceding example and other constants were $e = 0.001, f = 1.0$ and $g = 10.0$. The set of twelve values of d used in the preceding example was also used here and $k = 1, 2, 3$ were tried. Results corresponding to the minima of the ABIC's are illustrated in Fig.'s 2—4.

Fig. 2 shows the result of application of the present procedure to an artificial series given in Abe, Ito, Maruyama et al (1971, pp. 250-251). The result shows a very good reproduction of the true trend curve which was disturbed by a fixed multiplicative seasonality and the addition of the irregular components to produce the observations denoted by original.

It is remarkable that by this procedure no special treatment is necessary at the end of the series. This point is a significant advantage over the conventional procedures which require various ad hoc adjustments at the beginning and end of the series (Shiskin and Eisenpress, 1957). Fig. 3 shows the result of application to the last six years of the series of the logarithms of the number of airline passengers, given as Series *G* in Box and Jenkins (1970). The result reveals a very reasonable gradual change of the seasonality. The procedure has also been applied to the time series of labor force given in Table

1 of Shiskin and Eisenpress (1957, p.442) and the result is given in Fig. 4. The adjusted series is simply defined by $y(i) - S_i$ and is compared with the series adjusted by the Method II by Shiskin and Eisenpress.

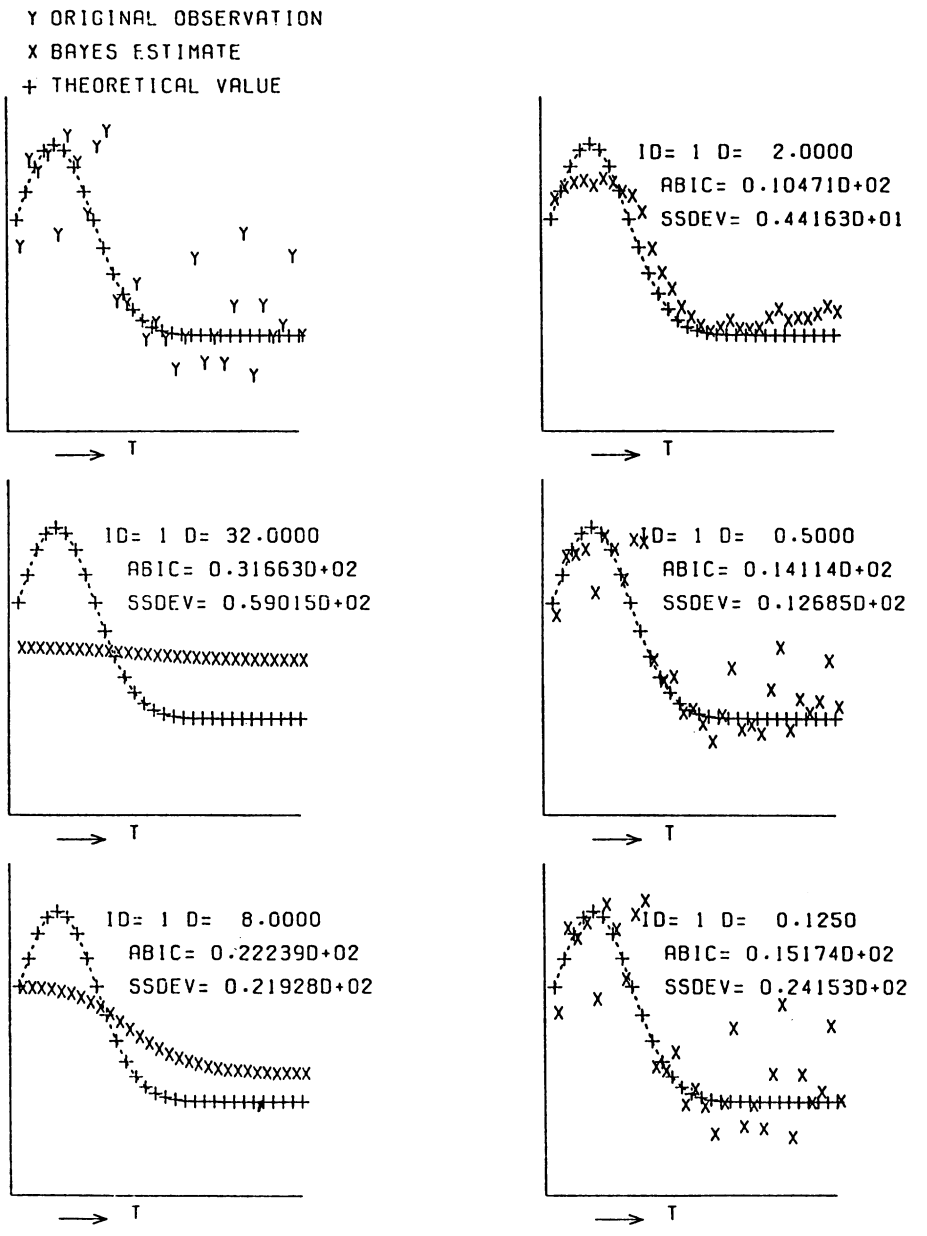


FIGURE 1

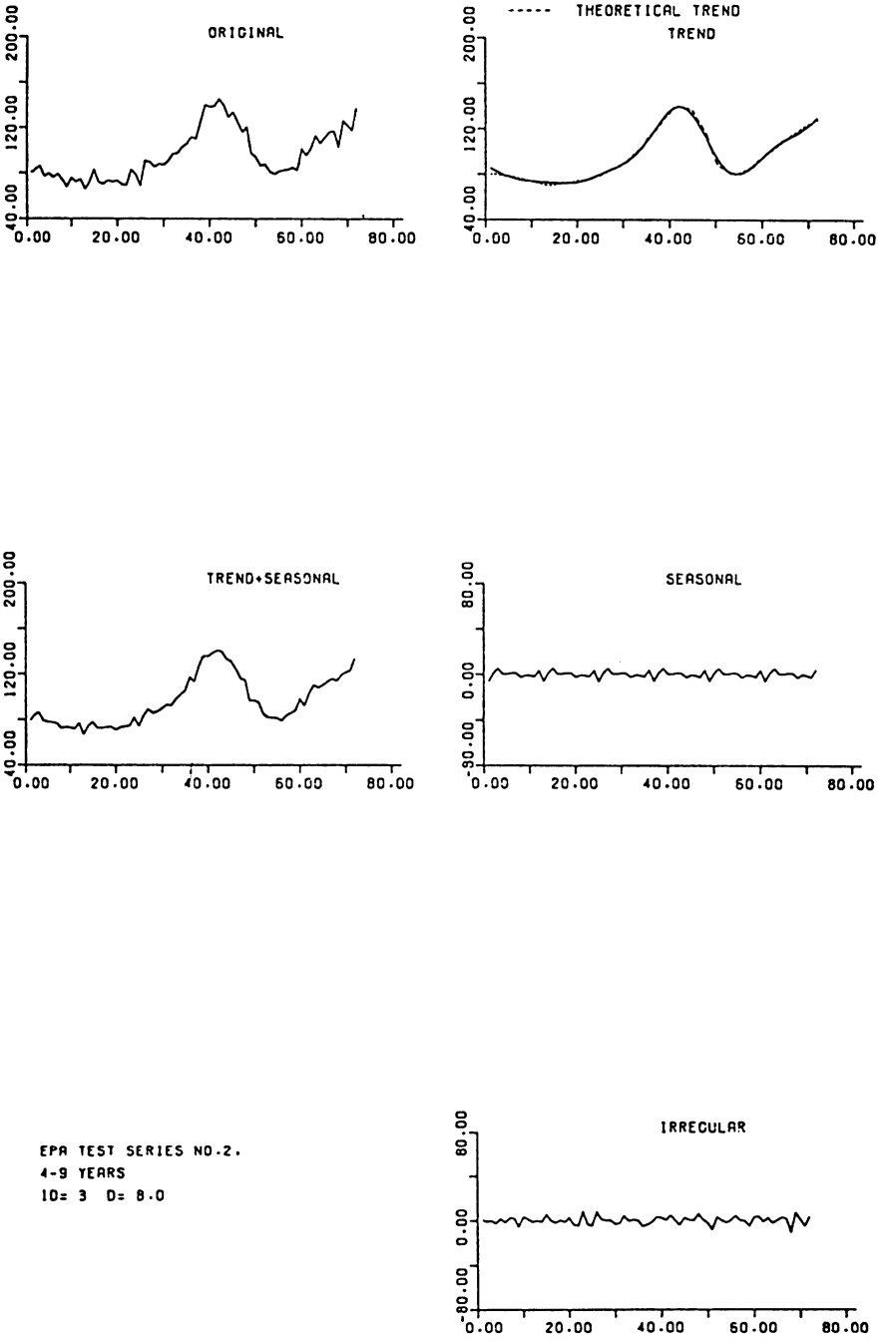

ORIGINAL
200.00
120.00
40.00
0.00
20.00
40.00
60.00
80.00
----- THEORETICAL TREND
TREND
TREND+SEASONAL
SEASONAL
80.00
0.00
-80.00
IRREGULAR
EPA TEST SERIES NO.2.
4-9 YEARS
ID= 3 D= 8.0

FIGURE 2

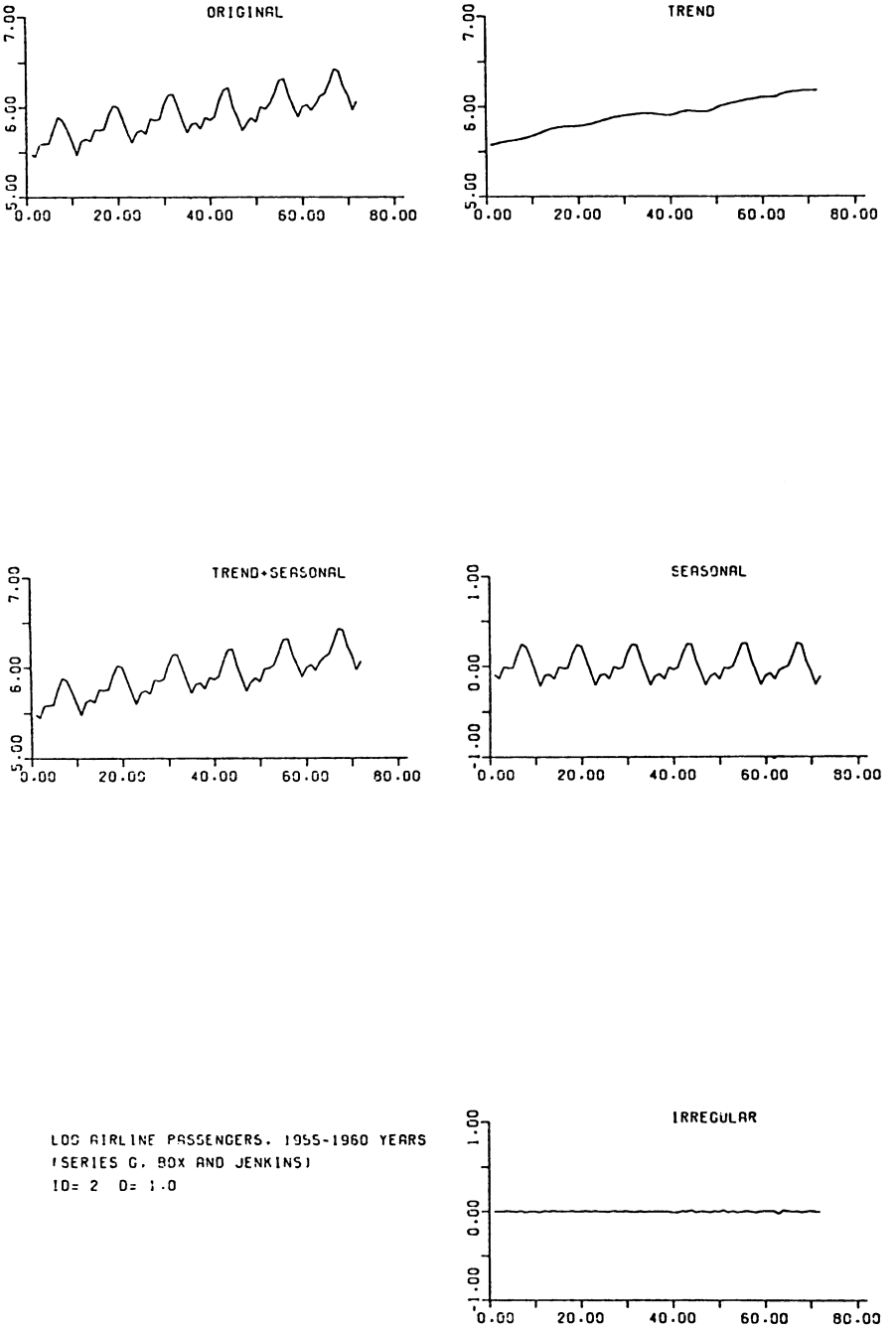

ORIGINAL
TREND
TREND+SEASONAL
SEASONAL
IRREGULAR
LOG AIRLINE PASSENGERS. 1955-1960 YEARS
(SERIES G. BOX AND JENKINS)
ID= 2 D= 1.0
7.00
6.00
5.00
1.00
0.00
-1.00
0.00
20.00
40.00
60.00
80.00

FIGURE 3

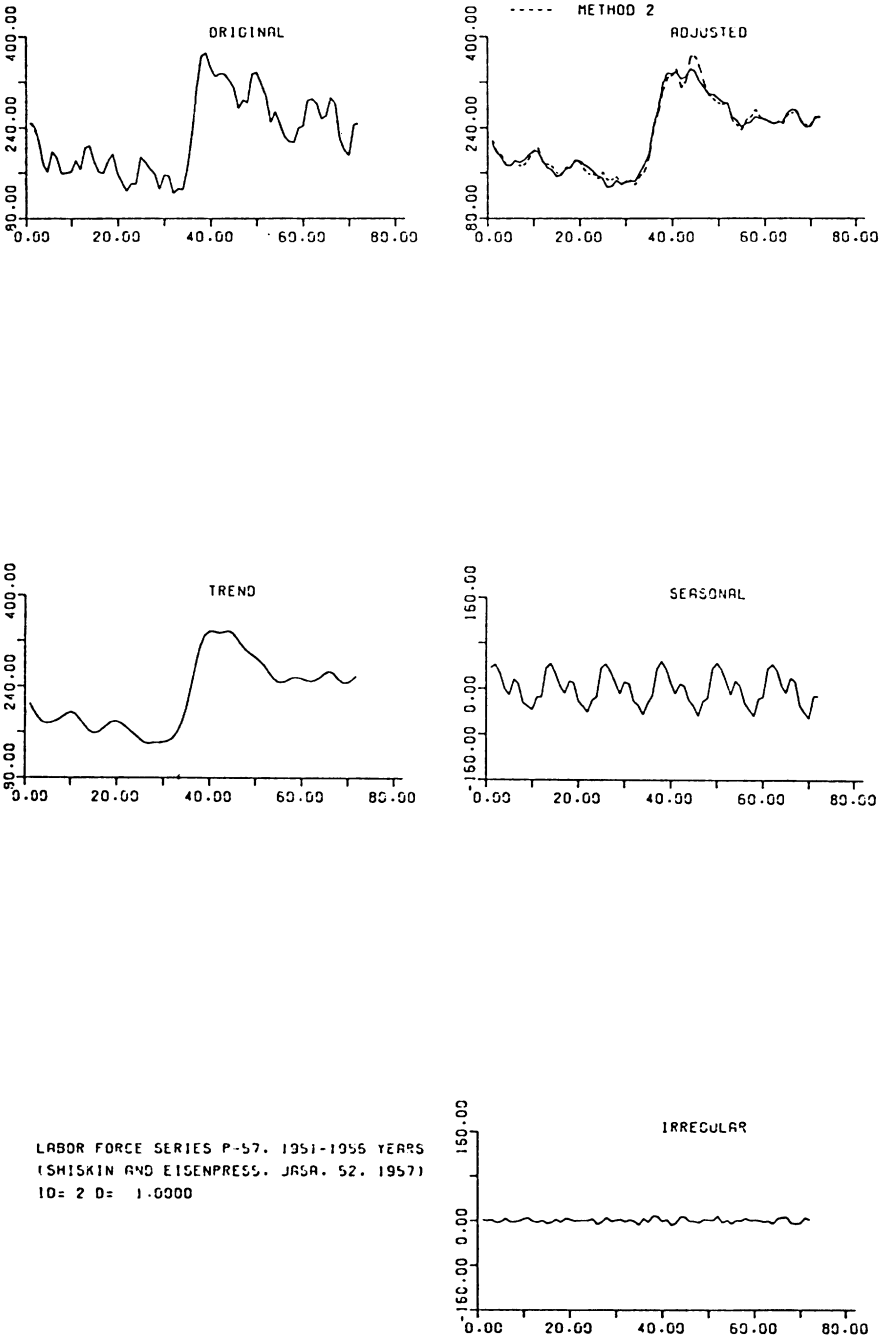

ORIGINAL
----- METHOD 2
ADJUSTED
TREND
SEASONAL
IRREGULAR
LABOR FORCE SERIES P-57. 1951-1956 YEARS
(SHISKIN AND EISENPRESS. JASA. 52. 1957)
ID= 2 D= 1.0000
400.00
240.00
80.00
160.00
0.00
-160.00
0.00
20.00
40.00
60.00
80.00

FIGURE 4

d. *Polynomial fitting*

By this example we wish to demonstrate that a reasonable definition of the likelihood of a model defined by the maximum likelihood estimates of the parameters can be given by exp (-(1/2) AIC) (Akaike, 1979a, c). The observations $y(i)$ are identical to those of the example *b* of this section and the polynomials of successively increasing order were fitted up to the 10th order by the method of maximum likelihood. Under the assumption of the Gaussian distribution, the AIC of the M^{th} order model is defined by

$$\text{AIC}\ (M) = N \log [\ (1/N)\ \text{S}(M)\ ] + 2M,$$

where $S(M)$ denotes the sum of squares of the residuals. Some of the estimated regression curves and the values of the AIC are illustrated in Fig. 5.

We smoothed these regression curves with the weight proportional to exp [- (1/2) AIC (M)] $\pi(M)$ with $\pi(M) \propto (M + 1)^{-1}$. The result is denoted by "Bayes" in the figure. The same type of procedure has been applied to the fitting of autoregressive models by Akaike (1979a) where the choice of $\pi(M)$ is discussed.

The present result shows that the procedure is practically useful, although its performance depends on the choice of the system of the basic functions or the polynomials. Usually this choice produces significant effects at the beginning and end of the regression curve. This shows the advantage of the models used in the preceding examples *b* and *c* over the present model. Nevertheless the present result demonstrates the feasibility of a Bayesian modeling of a multi-model problem with models defined with different number of parameters.

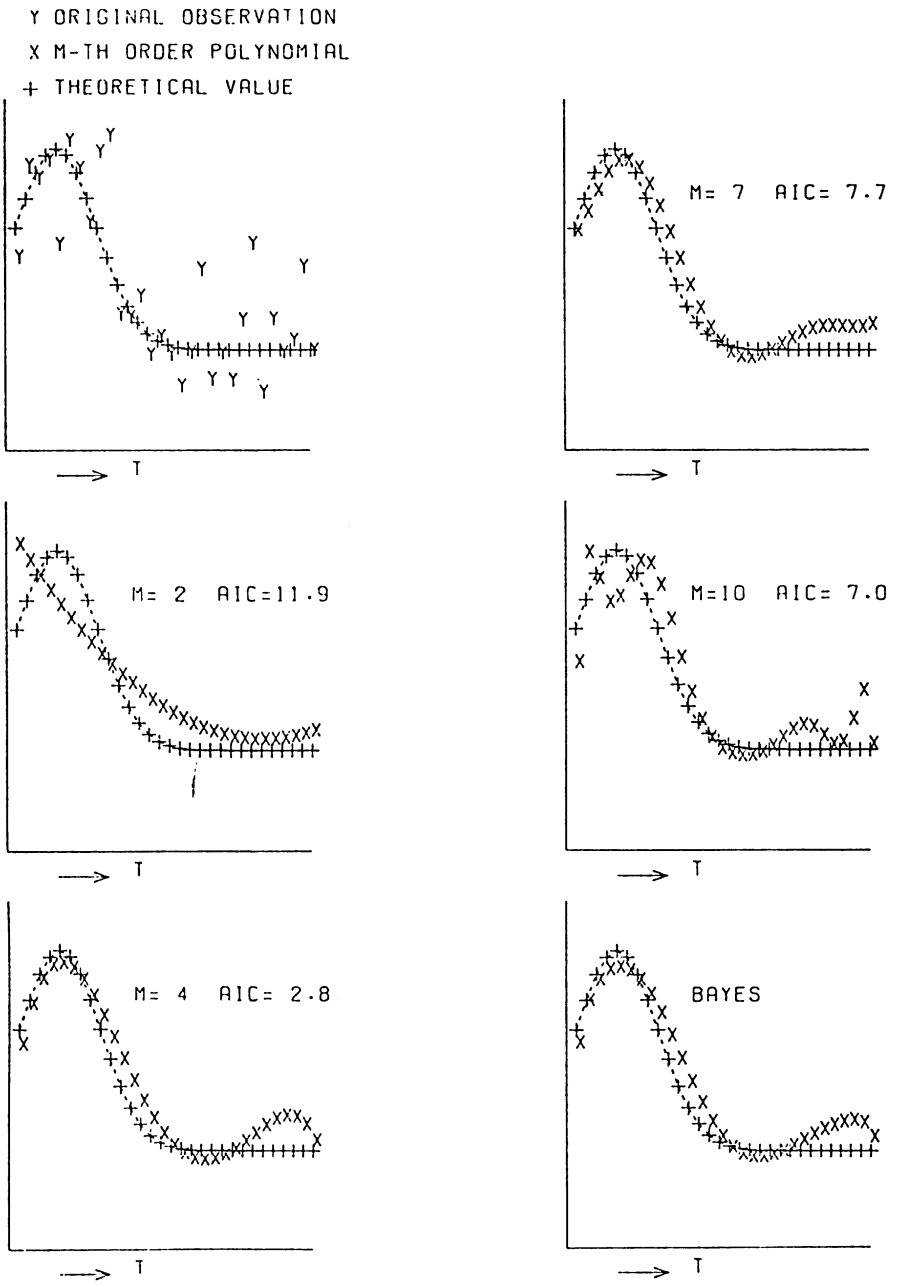

Y ORIGINAL OBSERVATION
X M-TH ORDER POLYNOMIAL
+ THEORETICAL VALUE
T
M= 7 AIC= 7.7
M= 2 AIC=11.9
M=10 AIC= 7.0
M= 4 AIC= 2.8
BAYES

FIGURE 5

6. DISCUSSION

The numerical results presented in the preceding section suggest the possibility of developing further applications of the general linear model to problems such as the gradually changing autoregression and the general trend analysis of time series. This possibility is pursued in Akaike (1979d). By choosing the set of d's properly the type II maximum likelihood method may be replaced by a procedure which takes an average of the models with respect to the weight proportional to the likelihood of each model. The performance of these procedures are controlled by the statistical characteristics of the related likelihood functions. One particular possibility is the extension of the concept of ignorance prior distribution to the prior distribution of a hyperparameter. This is discussed in Akaike (1980).

The application to seasonal adjustement is particularly interesting as it provides an example of the model which cannot be treated by the ordinary method of maximum likelihood. This example clearly demonstrates the practical utility of the Bayesian approach. It also shows that our present procedure may be characterized as a tempered method of maximum likelihood. The practical utility of the general linear model stems from the understandability and manipulability of the related prior distributions. This allows us to make proper judgement on how to temper the likelihood function through the choice of the values of the constants within the priors.

The subjective theory of probability is developed on the basis of our psychological reaction to uncertainty. Acordingly the final justification of the theory must be sought in the psychological satisfaction it can produce throught its application to real problems. It is only the accumulation of successful results of application that can really make the Bayesian statistics attractive.

The Bayes procedure provides a natural and systematic way of utilizing the information supplied by a likelihood function. The likelihood has a clearly defined objective meaning as the measure of the goodness of a model. It is this objectivity that provides the basis for the use of the subjective theory of probability as a guide in developing statistical procedures. Only this objectivity allows us to develop our confidence on the practical utility of the Bayes procedure, even when we know that the related model is our subjective construction.

ACKNOWLEDGEMENTS

The author is grateful to Dr K. Tanabe for the stimulating discussion of the linear model. The author is also indebted to Ms. E. Arahata for preparing the numerical and graphical outputs reported in this paper.

REFERENCES

ABE, K; ITO, M., MARUYAMA, A., YOSHIKAWA, J., ISUKADA, K. and IKEGAMI, M. (1971). *Methods of Seasonal Adjustements. Research Series No. 22.*, Tokyo: Economic Planning Agency Economic Research Institute (In Japanese).

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control,* **AC-19**, 716-723.

— (1978a). A new look at the Bayes procedure. *Biometrika,* **65**, 53-59.

— (1978b). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.,* **30**, *A*, 9-14.

— (1979a). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika,* **66**, 53-59.

— (1979b). A subjective view of the Bayes procedure. *Research Memo. No. 117.* Tokyo: The Institute of Statistical Mathematics. Revised, February 1979.

— (1979c). On the use of the predictive likelihood of a Gaussian model. *Research Memo. No 159.* Tokyo: The Institute of Statistical Mathematics.

— (1979d). On the construction of composite time series models. *Research Memo. No 161.* Tokyo: The Institute of Statistical Mathematics.

— (1980). Ignorance prior distribution of a hyperparameter and Stein's estimator. *Ann. Inst. Statist. Math.,* **32**, *A*, 171-178.

BOX, G.E.P. and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control.* San Francisco: Holden Day.

DE FINETTI, B (1974a). Bayesianism: Its unifying role for both the foundation and application of statistics. *Int. Stat. Rev.,* **42**, 117-130.

— (1974b/1975) *The Theory of Probability, Volumes 1 and 2.* New York: Wiley.

GOOD, I.J. (1965) *The Estimation of Probabilities.* Cambridge, Massachusetts: M.I.T. Press.

KUDO, H. (1973). The duality of parameter and sample. *Proceedings of the Institute of Statistical Mathematics Symposium,* **6**, 9-15 (In Japanese).

LINDLEY, D.V. (1957). A statistical paradox. *Biometrika,* **44**, 187-192.

O'HAGAN, A. (1978). Curve fitting and optimal design for prediction. *J.R. Statist. Soc. B,* **40**, 1-42.

SAVAGE, L.J. (1962). Subjective probability and statistical practice. In *The Foundations of Statistical Inference.* (G.A. Barnard and D.R. Cox eds.) 9-35.London: Methuen.

— (1954) *The Foundations of Statistics.* New York: Wiley.

SCHWARZ, G. (1978) Estimating the dimension of a model. *Ann. Statist.,* **6**, 461-464.

SHILLER, R. (1973) A distributed lag estimator derived from smoothness priors. *Econometrica,* **41**, 775-778.

SHISKIN, J. and EISENPRESS, H. (1957). Seasonal adjustments by electronic computer methods. *J. Amer. Statist. Ass.,* **52**, 415-499.

TIHONOV, A.N. (1965). Incorrect problems of linear algebra and a stable method for their solution. *Soviet Math. Dokl.,* **6**, 988-991.

WOLFOWITZ, J. (1962). Bayesian inference and axioms of consistent decision. *Econometrica,* **30**, 470-479.