Ann. Inst. Statist. Math.
32 (1980), Part A, 311-324

ON THE USE OF THE PREDICTIVE LIKELIHOOD OF A GAUSSIAN MODEL

HIROTUGU AKAIKE

(Received Sept. 3, 1979)

Abstract

The predictive likelihood of a model specified by data is defined when the model satisfies certain conditions. It reduces to the conventional definition when the model is specified independently of the data. The definition is applied to some Gaussian models and a method of handling the improper uniform prior distributions is obtained for the Bayesian modeling of a multi-model situation where the submodels may have different numbers of parameters. The practical utility of the method is checked by a Monte Carlo experiment of some quasi-Bayesian procedures realized by using the predictive likelihoods.

1. Introduction

Consider the simple problem of polynomial regression where the observation y_t at time t is represented by

$$(1.1) \qquad y_t=\sum_{i=0}^{m} \theta_i\phi_i(t)+z_t\ , \qquad t=1, 2,\cdots, N\,,$$

where $\phi_i(t)$ denotes the ith order polynomial satisfying the relations

$$\sum_{t=1}^{N} \phi_i(t)\phi_j(t)=\begin{cases} 1 & \text{for } i=j\ , \\ 0 & \text{otherwise}\ , \end{cases}$$

and z_t's are mutually independent Gaussian random variables with zero mean and variance σ^2 which is assumed to be known. Usually there is uncertainty about the choice of m, the order of the model. We may then take a Bayesian approach by assuming a prior distribution $p_m({}_m\theta)$ of ${}_m\theta=(\theta_0, \theta_1,\cdots, \theta_m)$ conditional on m and a prior probability $\pi(m)$, for $m=0, 1,\cdots, M$. The practical difficulty in this approach is the choice of $p_m({}_m\theta)$'s ; see, for example, the discussion of Atkinson and Cox ([6], p. 348).

From the construction of the model (1.1) one might wish to assume improper uniform prior distributions for $p_m({}_m\theta)$'s. The difficulty with the use of these improper uniform prior distributions in this case is the arbitrariness of their relative heights. This point is discussed earlier by Halpern [10].

To clarify the problem, consider more generally a set of models defined by the data distributions $\{f_k(x|{}_k\theta);\ k=1,2,\cdots,M\}$, where ${}_k\theta$ denotes the (vector) parameter of the kth model. Assume a prior distribution $p_k({}_k\theta)\pi(k)$, where $\pi(k)$ is the prior probability of the kth model and $p_k({}_k\theta)$ is the prior distribution of ${}_k\theta$ under the assumption of the kth model. The Bayes procedure produces the following transformation when data x is observed:

$$(1.2)\qquad p_k({}_k\theta)\pi(k)\rightarrow p_k({}_k\theta|x)\pi(k|x)\ ,$$

where

$$p_k({}_k\theta|x)=p_k(x)^{-1}f_k(x|{}_k\theta)p_k({}_k\theta)\ ,\qquad p_k(x)=\int f_k(x|{}_k\theta)p_k({}_k\theta)d_k\theta\ ,$$

and

$$\pi(k|x)=K_x^{-1}p_k(x)\pi(k)\ ,\qquad K_x=\sum p_k(x)\pi(k)\ .$$

The conditional posterior distribution $p_k({}_k\theta|x)$ is proportional to the product of the likelihood $f_k(x|{}_k\theta)$ and the prior probability density $p_k({}_k\theta)$. The posterior probability $\pi(k|x)$ of the kth model is proportional to the product of $p_k(x)$ and the prior probability $\pi(k)$. By the obvious analogy we call $p_k(x)$ the likelihood of the model specified by $f_k(\cdot|{}_k\theta)$ and $p_k({}_k\theta)$.

The conditional posterior distribution $p_k({}_k\theta|x)$ can be defined by formally applying the Bayes procedure even with an improper prior distribution $p_k({}_k\theta)$, if only the integral $\int f_k(x|{}_k\theta)p_k({}_k\theta)d_k\theta$ is finite. In this case $p_k({}_k\theta)$ may be replaced by $C_kp_k({}_k\theta)$ without causing any change of $p_k({}_k\theta|x)$, where C_k is an arbitrary positive constant. However, to use the value of the preceding integral, with $p_k({}_k\theta)$ replaced by $C_kp_k({}_k\theta)$, as the likelihood of the model, we must explicitly specify the values of C_k $(k=1,2,\cdots,M)$. This point is discussed clearly in Atkinson ([5], pp. 40–41) who concludes that the use of improper priors for comparing models cannot be justified. The problem here is how to define a reasonable quantity to replace $p_k(x)$ in (1.2).

In this paper we introduce the concept of predictive likelihood of the model specified by $f_k(\cdot|{}_k\theta)$ and an inferential distribution $p_k({}_k\theta|x)$ which is a distribution over the space of parameters ${}_k\theta$ and is determined by data x. By replacing $p_k(x)$ in (1.2) with this predictive likelihood, we get a quasi-Bayes procedure which can be applicable with improper prior distributions $p_k({}_k\theta)$, or even without these prior distributions, if the inferential distribution $p_k({}_k\theta|k)$ is appropriately defined. The practical utility of the quasi-Bayes procedure in its application to the estimation of the mean vector of a multivariate normal distribution is checked experimentally.

2. Predictive likelihood of a model

The point of view to consider the expected log likelihood of an assumed distribution, with respect to the true distribution of data, as a fundamental measure of the goodness of fit of the statistical model is developed in Akaike [1], [2]. From this point of view, the usefulness of a likelihood in statistical inference is due to the fact that the log likelihood is a natural estimate of the expected log likelihood of the assumed distribution. The likelihood $p_k(x)$, defined in Introduction, of the model specified by $f_k(\cdot|{}_k\theta)$ and $p_k({}_k\theta)$ is the conventional likelihood of the marginal distribution $p_k(\cdot)=\int f_k(\cdot|{}_k\theta)p_k({}_k\theta)d_k\theta$, or of k, with respect to the present data x and provides a measure of the goodness of $p_k(\cdot)$ as an approximation to the true distribution of the data x. The cause of the difficulty of an improper prior distribution discussed in Introduction is now clear. When $p_k({}_k\theta)$ is improper, $f_k(\cdot|{}_k\theta)$ and $p_k({}_k\theta)$ do not really define any statistical model to determine a distribution $p_k(\cdot)$. There can be no likelihood when there is no distribution.

Here we consider that the data x is taken from a fixed "true" distribution $f(\cdot)$. Imagine that a future observation y is also taken from $f(\cdot)$, independently of x. The future likelihood of the model specified by $f_k(\cdot|{}_k\theta)$ and an inferential distribution $p_k({}_k\theta|x)$, is then defined by

$$p_k(y|x)=\int f_k(y|{}_k\theta)p_k({}_k\theta|x)d_k\theta\ . \tag{2.1}$$

When $p_k(y|x)$ is considered as the distribution $p_k(\cdot|x)$ of future observation y it is called the predictive distribution. Thus the future likelihood is the likelihood of the predictive distribution with respect to a particular y. The goodness of the model as an approximation to $f(\cdot)$ is measured by the expected log future likelihood $E_y \log p_k(y|x)$, where E_y denotes the expectation with respect to the true distribution $f(\cdot)$ of y.

When y is available, $\log p_k(y|x)$ forms a natural estimate of $E_y \log p_k(y|x)$. Actually we do not have y and consider the use of $\log p_k(x|x)$ in place of $\log p_k(y|x)$ as an "estimate" of the expected log future likelihood. When the true distribution $f(\cdot)$ is not approximated well by $f_k(\cdot|{}_k\theta)$ for any choice of k and ${}_k\theta$, the situation is uninteresting. Thus we assume that $f(\cdot)$ is a member of some of the families $\{f_k(\cdot|{}_k\theta)\}$ ($k=1,2,\cdots$). When $f(\cdot)$ is not a member of $\{f_k(\cdot|{}_k\theta)\}$, k fixed, $E_y \log p_k(y|x)$ and $\log p_k(x|x)$ will generally be significantly lower compared with the case where $f(\cdot)$ is a member. Thus we pay our attention to the statistical characteristic of $\log p_k(x|x)$ when $f(\cdot)$ is a member of $\{f_k(\cdot|{}_k\theta)\}$. We further restrict our attention to those inferential distribution $p_k({}_k\theta|x)$ for which, when $f(\cdot)$ is a member of $\{f_k(\cdot|{}_k\theta)\}$, i.e., under the assumption of the kth model, we have

$$E_x[E_y \log p_k(y|x) - \log p_k(x|x)] = c_k\,, \tag{2.2}$$

where c_k is a constant. We try to correct this bias, or the expected deviation of $\log p_k(x|x)$ from $E_y \log p_k(y|x)$, by adding c_k to $\log p_k(x|x)$ and define the log predictive likelihood of the model defined by $f_k(\cdot|{}_k\theta)$ and $p_k({}_k\theta|x)$ by

$$l[f_k(\cdot|{}_k\theta),\, p_k({}_k\theta|x)] = \log p_k(x|x) + c_k\,. \tag{2.3}$$

The predictive likelihood of the model is then defined by

$$L_k(x) = \exp\{l[f_k(\cdot|{}_k\theta),\, p_k({}_k\theta|x)]\}\,. \tag{2.4}$$

Certainly the above bias correction is meaningful only when $f(\cdot)$ is a member of the kth family. The use of the present definition thus depends on the assumption that the expected amount of decrease of $\log p_k(x|x)$ is larger than the correction by c_k, when $f(\cdot)$ is not a member of the family.

A referee pointed out that $\log p_k(y^*|x)$ could be used as an estimate of $\log p_k(y|x)$, where y^* is any prediction of y based on the data x at hand. This is certainly an interesting suggestion which deserves further analysis.

3. Predictive likelihood of a Gaussian model

Here we will show that predictive likelihoods can be defined for a rather general class of Gaussian models. Consider the case where $f_k(\cdot|{}_k\theta)$ is an L-dimensional Gaussian distribution $N({}_k\theta, \Sigma_1)$ and the inferential distribution $p({}_k\theta|x)$ is $N(\theta(x), \Sigma_2)$, with the mean $\theta(x)$ defined as a function of the data x. Here Σ_1 and Σ_2 are assumed to be known. The predictive distribution $p_k(\cdot|x)$ is then $N(\theta(x), \Sigma_1+\Sigma_2)$ and for $f(\cdot) = f_k(\cdot|{}_k\theta_0)$ we have

$$\begin{aligned} &E_x E_y(-2)\log p_k(y|x) \\ &\quad = \mathrm{tr}\,(\Sigma_1+\Sigma_2)^{-1}[\Sigma_1 + E_x(\theta(x)-{}_k\theta_0)(\theta(x)-{}_k\theta_0)'] + C_k \end{aligned} \tag{3.1}$$

and

$$(3.2)\qquad E_x(-2)\log p_k(x\,|\,x)=\operatorname{tr}(\Sigma_1+\Sigma_2)^{-1}E_x(x-\theta(x))(x-\theta(x))'+C_k\ ,$$

where C_k denotes a common constant independent of ${}_k\theta_0$ and E_x the expectation with respect to the distribution $f(x)=f_k(x\,|\,{}_k\theta_0)$ of x. Since $E_x x={}_k\theta_0$, the difference of (3.1) and (3.2) is constant, irrespectively of the choice of ${}_k\theta_0$, if the covariance matrices of $x-\theta(x)$ and $\theta(x)-{}_k\theta_0$ are independent of ${}_k\theta_0$. This is the case when $\theta(x)-E_x\theta(x)$ is a function of $x-{}_k\theta_0$ only. This condition is satisfied for $\theta(x)=Kx$ defined with a constant $L\times L$ matrix K. For this case we have

$$(3.3)\qquad \begin{aligned} c_k&=E_x[E_y\log p_k(y\,|\,x)-\log p_k(x\,|\,x)]\\ &=\left(-\frac{1}{2}\right)\operatorname{tr}(\Sigma_1+\Sigma_2)^{-1}[K\Sigma_1+\Sigma_1K']\ , \end{aligned}$$

and the log predictive likelihood is given by

$$(3.4)\qquad \begin{aligned} l_k(x)=\left(-\frac{1}{2}\right)&[L\log 2\pi+\log|\Sigma_1+\Sigma_2|\\ &+\operatorname{tr}(\Sigma_1+\Sigma_2)^{-1}(I-K)xx'(I-K)']+c_k\ , \end{aligned}$$

where $|\Sigma|$ denotes the determinant of Σ. The examples to be discussed in the following are of this type, including the cases which can be defined as the limits when some of the diagonal elements of Σ_2 tend to zero.

As a concrete example, consider the situation of estimation of the mean of a multivariate normal distribution. Take the true distribution $f(\cdot)$ as an L-dimensional normal distribution $N(\theta_0, I)$ with mean $\theta_0=(\theta_{01}, \theta_{02}, \cdots, \theta_{0L})$ and variance matrix I, an $L\times L$ identity matrix. We consider the use of a set of models specified by

$$f_k(x\,|\,\theta)=\left(\frac{1}{2\pi}\right)^{L/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{L}(x_i-\theta_i)^2\right]\ ,\qquad k=0, 1, \cdots, 2^L-1\ ,$$

where $\theta_i=0$, when the ith bit of the binary representation of k is 0, unconstrained, otherwise. For the sake of simplicity of presentation we consider the case where $k=2^m-1$, i.e., the case where only the first $m\theta_i$'s are retained and others are put equal to 0, and represent the model by

$$(3.5)\qquad \begin{aligned} f_k(x\,|\,{}_k\theta)=\left(\frac{1}{2\pi}\right)^{m/2}&\exp\left[-\frac{1}{2}\sum_{i=1}^{m}(x_i-{}_k\theta_i)^2\right]\left(\frac{1}{2\pi}\right)^{(L-m)/2}\\ &\cdot\exp\left(-\frac{1}{2}\sum_{i=m+1}^{L}x_i^2\right)\ , \end{aligned}$$

where ${}_k\theta=({}_k\theta_1, {}_k\theta_2, \cdots, {}_k\theta_m)$. Any other model can be brought into this form by a proper reordering of the suffices i. Assuming an improper uniform prior distribution as $p_k({}_k\theta)$ and applying the Bayes procedure formally, we get

$$p_k({}_k\theta\,|\,x)=\left(\frac{1}{2\pi}\right)^{m/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}({}_k\theta_i-x_i)^2\right] \tag{3.6}$$

as our posterior distribution of ${}_k\theta$. From (2.1), (3.5) and (3.6) we have

$$p_k(y\,|\,x)=\left(\frac{1}{2\pi}\right)^{L/2}\left(\frac{1}{2}\right)^{m/2}\exp\left[-\frac{1}{4}\sum_{i=1}^{m}(y_i-x_i)^2-\frac{1}{2}\sum_{i=m+1}^{L}y_i^2\right]. \tag{3.7}$$

Thus we get

$$E_xE_y\log p_k(y\,|\,x)=-\frac{m}{2}\log 2-\frac{m}{2}-\frac{1}{2}(L-m)+D$$

and

$$E_x\log p_k(x\,|\,x)=-\frac{m}{2}\log 2-\frac{1}{2}(L-m)+D\,,$$

where E_x denotes the expectation under the assumption that $f(\cdot)$ is a member of the family $\{f_k(\cdot\,|\,{}_k\theta)\}$ and $D=-(1/2)L\log 2\pi$. The c_k required for the bias correction is then given by

$$c_k=-\frac{m}{2}\,,$$

where m is the number of unconstrained θ_i's in the kth model. This result could have been obtained directly by (3.3). The log predictive likelihood of the model is given by $l_k(x)=\log p_k(x\,|\,x)+c_k$ and, with (3.7), the predictive likelihood is given by

$$L_k(x)=C\exp\left[-\frac{1}{2}(S_L-S_m+m\log 2+m)\right], \tag{3.8}$$

where S_m denotes the sum of squares of x_i for which $\theta_i=0$ is not assumed and C is a positive constant.

We could have started with $p_k({}_k\theta\,|\,x)=\delta({}_k\theta-{}_kx)$, where ${}_kx$ is the maximum likelihood estimate of ${}_k\theta$ and is a vector composed of the components of x corresponding to those of ${}_k\theta$. With this choice of $p_k({}_k\theta\,|\,x)$ we have $p_k(y\,|\,x)=f_k(y\,|\,{}_kx)$ and analogously to (3.8) we get

$$L_k(x)=C\exp\left[-\frac{1}{2}\left(\sum_{i=m+1}^{L}x_i^2+2m\right)\right]. \tag{3.9}$$

The quantity within the parentheses on the right side is essentially the AIC (an information criterion) statistic (Akaike [1]) which is defined by

$$\mathrm{AIC}\,(k)=(-2)\log\max_{{}_k\theta} f_k(x\,|\,{}_k\theta)+2m\ ,$$

and we get $L_k(x)=C\exp[-\mathrm{AIC}\,(k)/2]$. The minimum AIC estimate of θ_0 is defined by ${}_kx$ with k for which AIC (k) is minimum, i.e., for which $L_k(x)$ is maximum. For later use we will denote this estimate by θ_{IC}. Here IC stands for "information criterion".

4. Discussion

When we assume $p_k({}_k\theta\,|\,x)=\delta({}_k\theta-{}_k\theta_0)$, the distribution concentrated at one particular parameter value ${}_k\theta_0$, we have $p_k(\cdot\,|\,x)=f_k(\cdot\,|\,{}_k\theta_0)$. Since that $E_x\log p_k(x\,|\,x)=E_x\log f_k(x\,|\,{}_k\theta_0)=E_y\log f_k(y\,|\,{}_k\theta_0)=E_y\log p_k(y\,|\,x)$ holds, we get $L_k(x)=\exp[\log p_k(x\,|\,x)]=f_k(x\,|\,{}_k\theta_0)$, the conventional likelihood of the model specified by $f_k(\cdot\,|\,{}_k\theta_0)$. More generally, if we assume $p_k({}_k\theta\,|\,x)$ $=p_k({}_k\theta)$ a proper prior distribution independent of data, we get $L_k(x)=\exp$ $\left[\log\int f_k(x\,|\,{}_k\theta)p_k({}_k\theta)d_k\theta\right]$, which is the likelihood $p_k(x)$ of the model specified by $f_k(\cdot\,|\,{}_k\theta)$ and $p_k({}_k\theta)$, as defined in Introduction. These results suggest that the present definition of the likelihood $L_k(x)$ is a natural extension of the classical likelihood to the present particular models.

One major motivation for the introduction of the present definition of the likelihood of a model was to provide a reasonable procedure of the use of improper uniform prior distributions of the parameters with different dimensionalities. One may hope that the difficulty of the relative heights of the improper prior distributions may be solved by considering a limit of some properly chosen proper prior distributions. To show that it is not easy to realize the idea, we first assume the proper prior distributions of the parameters given by

$$(4.1)\qquad p_k({}_k\theta)=\left(\frac{1}{2\pi\tau^2}\right)^{m/2}\exp\left[-\frac{1}{2\tau^2}\sum_{i=1}^{m}{}_k\theta_i^2\right]\ ,$$

where τ^2 is a constant common to all $p_k({}_k\theta)$'s. It is easy to check that the likelihood of the model specified by (3.5) and (4.1) is given by

$$(4.2)\qquad p_k(x)=\left(\frac{1}{2\pi}\right)^{L/2}\exp\left[-\frac{1}{2}\left(S_L-S_m+\frac{1}{1+\tau^2}S_m+m\log(1+\tau^2)\right)\right]\ ,$$

where S_L denotes the sum of squares of x_i $(i=1,2,\cdots,L)$ and S_m the sum of squares of x_i for which $\theta_i=0$ is not assumed. As we increase τ^2 indefinitely the ratio of the likelihood of a model with $m\neq 0$ to that of the model with $m=0$ goes down to zero. This means that, by this passage to the uniform prior distributions, only the simplest model, with all the θ_i's equal to 0, is retained. In the case of the polynomial regression, discussed in Introduction, this means that only the 0th order model is possible. Although this result is somewhat unexpected, it is natural because it is only the model with all the θ_i's assumed to be zeros that can produce finite x_i's when τ^2 is infinitely large. Thus this approach does not lead to any sensible solution of the difficulty.

If, instead of $p_k({}_k\theta)$, we start with the conditional posterior distribution

$$(4.3)\qquad p_k({}_k\theta\,|\,x)=\left(\frac{1}{2\pi}\,\frac{1+\tau^2}{\tau^2}\right)^{m/2}\exp\left[-\left(\frac{1+\tau^2}{2\tau^2}\right)\sum_{i=1}^{m}\left({}_k\theta_i-\frac{\tau^2}{1+\tau^2}x_i\right)^2\right],$$

the model satisfies the condition which leads to (3.3) and, from (3.4), the predictive likelihood of the kth model is obtained as

$$(4.4)\qquad L_k(x)=C\exp\left[-\frac{1}{2}\left(S_L-S_m+\frac{S_m}{(1+\tau^2)(1+2\tau^2)}\right.\right.$$
$$\left.\left.+m\log\left(\frac{1+2\tau^2}{1+\tau^2}\right)+m\left(\frac{2\tau^2}{1+2\tau^2}\right)\right)\right].$$

As τ^2 is increased indefinitely this tends to $L_k(x)$ of (3.8), the predictive likelihood of a model obtained by assuming the uniform improper prior distribution of ${}_k\theta$. This shows that it is the present definition (4.4) and not (4.2) that is in consonance with the conventional definition of the improper uniform prior distribution as the limit of $p_k({}_k\theta)$ (4.1) when τ^2 tends to infinity.

When the predictive likelihood $L_k(x)$ is available, the above results suggest the replacement of the exact Bayes procedure, schematically represented by

Procedure I (exact Bayes procedure)

$$\begin{tikzcd}
 & p_k({}_k\theta\,|\,x) \arrow[dr, dash] & \\
p_k({}_k\theta) \arrow[ur] \arrow[dr] & & p_k({}_k\theta\,|\,x)p_k(x)\,, \\
 & p_k(x) \arrow[ur] &
\end{tikzcd}$$

with a quasi-Bayes procedure represented by

Procedure II (quasi-Bayes procedure)

$$p_k({}_k\theta)\longrightarrow p_k({}_k\theta\,|\,x)\longrightarrow L_k(x)\longrightarrow p_k({}_k\theta\,|\,x)L_k(x)\,.$$

It was our observation that only Procedure II could produce meaningful result when the prior distribution $p_k({}_k\theta)$ tended to the improper uniform prior distribution. The basic ingredients of Procedure II are $f_k(x\,|\,{}_k\theta)$ and $p_k({}_k\theta\,|\,x)$ and we could have started with an inferential distribution $p_k({}_k\theta\,|\,x)$ which is not necessarily a posterior distribution based on a proper prior distribution. Thus we will denote by Procedure II that part of the transformation which transforms $p_k({}_k\theta|x)$ into $p_k({}_k\theta|x)\cdot L_k(x)$. In the formal Bayes procedure, the posterior distribution $p_k({}_k\theta|x)$ is invariant with the multiplication of the prior density $p_k({}_k\theta)$ by a positive constant. Thus the problem of the relative heights of improper prior distributions, discussed in Introduction, disappears with Procedure II. This is due to the fact that the procedure weights $p_k({}_k\theta|x)$ by evaluating $p_k({}_k\theta|x)$ by $f_k(x|{}_k\theta)$, while Procedure I does it through the evaluation of $p_k({}_k\theta)$ by $f_k(x|{}_k\theta)$. The only question is now whether Procedure II can produce practically useful results, and this must be checked with real applications.

5. Applications

To demonstrate the practical utility of Procedure II of the preceding section, here it is applied to construct some quasi-Bayes procedures for the problem of estimation of the mean of a multivariate normal distribution using the models of Section 3.

5.1. *Construction of a Bayesian model with uniform* $p_k({}_k\theta)$

We start with the assumption of improper uniform prior distributions $p_k({}_k\theta)=\text{const}$. To incorporate the possibility of nearly zero means we consider the use of the prior probabilities of the models specified by $\pi(k|\rho)=\rho^m(1-\rho)^{L-m}$ with a hyperparameter ρ $(0<\rho<1)$. Here m denotes the dimension of ${}_k\theta$ and $1-\rho$ represents the prior probability of ${}_k\theta_i=0$. We further assume a uniform distribution of ρ over $(0, 1)$ and the prior probability $\pi(k)$ of the kth model is given by the integration of $\pi(k|\rho)$ over $0<\rho<1$. The "posterior" probability of the kth model is given by $\pi(k|x)=K_0^{-1}L_k(x)\pi(k)$, where $K_0=\sum L_k(x)\pi(k)$ and $L_k(x)$ is the likelihood defined by (3.4). The "posterior" distribution is defined by $p_k({}_k\theta|x)\pi(k|x)$, where $p_k({}_k\theta|x)$ is given by (3.2). The mean of this "posterior" distribution defines an estimate of θ_0 given by $\theta_{UF}=(\theta_{UF1}, \theta_{UF2}, \cdots, \theta_{UFL})$ with θ_{UFi} defined by

$$\theta_{UFi}=[\sum_{k\in I_i}\pi(k|x)]x_i\ , \qquad i=1, 2, \cdots, L\ ,$$

where I_i denotes the set of k's for which $\theta_i=0$ is not assumed. Here UF stands for "uniform $p_k({}_k\theta)$ and fixed $\pi(k)$".

To see the performance of θ_{UF} the sum of squared errors $\|\theta_{UF}-\theta_0\|^2=\sum(\theta_{UF}-\theta_{0i})^2$ has been evaluated by Monte Carlo experiments for θ_0's defined by

$$\theta_{0i}=Az_i\ , \qquad \text{if } 1\leqq i\leqq M\ ,$$

$$\theta_{0i}=\varepsilon z_i\ , \qquad \text{if } M+1\leqq i\leqq L\ ,$$

where $z=(z_1, z_2, \cdots, z_L)$ is a random sample from $N(0, I)$, A is a variable scaling constant and $\varepsilon=1/3$. The value $\varepsilon=1/3$ was chosen to represent "non-significant" parameters more realistically than $\varepsilon=0$. The results are summarized in Table 1, where, for the purpose of comparison, errors of the least-squares estimator θ_{LS} $(=x)$, the minimum AIC estimator θ_{IC}, the positive part James-Stein estimator θ_{JS} $(=[1-(L-2)\|x\|^{-2}]^{+}x)$ and the estimators θ_{UA} and θ_{AF}, which are to be defined in the following subsections, are also illustrated. It can be seen that the present estimator θ_{UF} shows good adaptation to the change of the distribution of θ_0 compared with θ_{JS}, yet without the increase of error induced by θ_{IC} for the case where $M=L$. This result demonstrates the power of the present quasi-Bayes procedure. The normal random numbers were generated simply by adding twelve random numbers uniformly distributed over [0, 1], which were generated by a multiplicative congruential method, and then subtracting 6 to keep the mean equal to 0. Similar experiments were conducted using a physical noise source

Table 1. Average sum of squared errors[a]

A	θ_{LS}	$\theta_{JS}(SD^b)$	θ_{IC}	θ_{UF}	θ_{UA}	$\theta_{AF}(SD^b)$
		a. $L=4$ $M=4$				
0.5	3.95	1.90(0.063)	2.87	1.85	1.98	1.23(0.044)
1.0	3.95	2.70(0.067)	3.99	2.87	2.97	2.35(0.057)
3.0	3.95	3.74(0.081)	4.74	4.04	4.06	4.12(0.089)
10.0	3.95	3.93(0.085)	4.22	3.97	3.97	4.13(0.090)
100.0	3.95	3.95(0.086)	3.98	3.95	3.95	4.05(0.089)
		b. $L=4$ $M=2$				
0.5	3.95	1.79(0.063)	2.73	1.71	1.85	1.09(0.043)
1.0	3.95	2.29(0.066)	3.27	2.28	2.39	1.70(0.051)
3.0	3.95	3.37(0.077)	3.64	3.19	3.20	2.91(0.076)
10.0	3.95	3.82(0.083)	3.37	3.24	3.25	2.82(0.076)
100.0	3.95	3.95(0.086)	3.21	3.27	3.28	2.35(0.072)
		c. $L=6$ $M=6$				
0.5	5.95	2.24(0.063)	4.35	2.70	2.93	1.78(0.049)
1.0	5.95	3.78(0.073)	6.00	4.40	4.47	3.56(0.067)
3.0	5.95	5.54(0.096)	7.14	5.99	6.06	6.00(0.104)
10.0	5.95	5.91(0.104)	6.39	5.97	5.97	6.15(0.107)
100.0	5.95	5.95(0.105)	6.00	5.95	5.95	6.08(0.108)
		d. $L=6$ $M=2$				
0.5	5.95	1.92(0.063)	4.00	2.37	2.65	1.47(0.048)
1.0	5.95	2.62(0.067)	4.55	3.03	3.21	2.13(0.055)
3.0	5.95	4.52(0.089)	4.92	4.21	4.12	3.55(0.082)
10.0	5.95	5.61(0.099)	4.64	4.39	4.20	3.37(0.083)
100.0	5.95	5.94(0.105)	4.49	4.46	4.24	2.61(0.074)

[a] Averaged over 1000 samples.

[b] Square root of (sample variance/1000).

to generate the random numbers. The results did not show any substantial deviations from the result given in Table 1.

5.2. *Data adaptive choice of $\pi(k)$*

Testing the performance of the quasi-Bayes procedure defined in the preceding subsection with $\pi(k)$ put equal to some particular $\pi(k|\rho)$ we noticed that the sensitivity of the procedure to the variation of ρ was rather low. This suggested that only if we could avoid a gross misspecification of the prior distribution would the corresponding quasi-Bayes procedure produce practically useful results. The validity of this idea was checked by a data adaptive choice of $\pi(k)$ realized by putting $\pi(k)=\pi(k|\rho_0)$, where ρ_0 is such that the expected log likelihood of $p_k({}_k\theta|x)\pi(k|\rho)$ with respect to the "posterior" distribution $Cp_k({}_k\theta|x)L_k(x)\pi(k)$ with $\pi(k)=1/L$ attains the maximum at $\rho=\rho_0$. Since the expected log likelihood is synomimous to the probabilistic definition of entropy (Akaike [1]), we call this type of method of fitting a distribution the method of maximum entropy. In the present case it reduces to maximizing $\sum L_k(x)\log\pi(k|\rho)$ with respect to ρ.

The quasi-Bayes procedure defined by replacing $\pi(k)$ of the procedure defined in the preceding subsection by $\pi(k|\rho_0)$ showed only minor deviation from the original. The mean of the "posterior" distribution obtained by this procedure is denoted by θ_{UA} and its performance as an estimator of θ_0 is also illustrated in Table 1. Here *UA* stands for "uniform $p_k({}_k\theta)$ and adaptive $\pi(k)$".

5.3. *Data adaptive choice of $p_k({}_k\theta)$*

Although θ_{UF} has shown reasonable adaptability to the change of the distribution of θ_0 one may wish to incorporate further prior information of the possibility of happening of rather insignificant values of the parameters by introducing some proper prior distribution of ${}_k\theta$. Due to the indefiniteness of the range of related quantities this task is not so simple as that of specifying the prior probabilities $\pi(k)$.

One possibility is to take $p_k({}_k\theta)$ of (4.1) and assume some improper prior distribution of τ^2. This leads to an improper prior distribution of ${}_k\theta$ which will replace the uniform prior distribution $p_k({}_k\theta)$ in the procedure of 5.1. Although not infeasible, this procedure requires further analytical manipulations of the related distributions.

The comparison of the performances of θ_{UF} and θ_{UA} introduced in the preceding subsections suggests that a procedure similar to that of the determination of $\pi(k|\rho_0)$ might be worth trying for the data adaptive choice of τ^2 of (4.1). To realize this we apply the method of maximum entropy to optimize the fit of $p_k({}_k\theta|\tau)\pi(k)$ to the "posterior" distribution $p_k({}_k\theta|x)\pi(k|x)$ of subsection 5.1, with $\pi(k)$ also given in that subsection. This reduces to the maximization of $\sum_k \pi(k|x)\int p_k({}_k\theta|x)\log p_k({}_k\theta|\tau)d_k\theta$ with respect to τ^2. Since we have

$$\int p_k({}_k\theta|x)\log p_k({}_k\theta|\tau)d_k\theta$$
$$=-\left(\frac{1}{2}\right)\left[m\log 2\pi+m\log\tau^2+\frac{1}{\tau^2}(m+S_m)\right],$$

the value of τ^2 that maximizes the entropy is given by $\tau_0^2=1+\bar{S}_m/\bar{m}$, where $\bar{S}_m=\sum S_m\pi(k|x)$ and $\bar{m}=\sum m\pi(k|x)$. Here the summation is over k and m denotes the dimension of ${}_k\theta$.

By using $p_k({}_k\theta|\tau_0^2)$ formally as the prior distribution, the "likelihood" of the kth model and the conditional "posterior" distribution are respectively given by (4.2) and (4.3) with $\tau^2=\tau_0^2$. Our estimate of θ_0 is defined as the mean of the "posterior" distribution and is given by $\theta_{AF}=(\theta_{AF1},\theta_{AF2},\cdots,\theta_{AFL})$ with θ_{AFi} defined by

$$\theta_{AFi}=\left(\frac{\tau_0^2}{1+\tau_0^2}\right)\left[\sum_{k\in I_i}\pi(k|x;\tau_0)\right]x_i\,,\qquad i=1,2,\cdots,L\,,$$

where $\pi(k|x;\tau_0)=p_k(x;\tau_0)\pi(k)$ and I_i denotes the set of k's for which $\theta_i=0$ is not assumed. Here AF stands for "adaptive $p_k({}_k\theta)$ and fixed $\pi(k)$".

The performance of this estimator θ_{AF} can be evaluated by the result of the Monte Carlo experiment given in Table 1. A better adaptation to the change of the distribution of θ_0 than that of θ_{JS} or θ_{UF} is obtained. The penalty for this improved adaptability is the slight

Table 2. Examples of estimates

First example						Second example					
θ_0	x	θ_{JS}	θ_{UF}	θ_{UA}	θ_{AF}	θ_0	x	θ_{JS}	θ_{UF}	θ_{UA}	θ_{AF}
0	−.32	−.26	−.19	−.13	−.13	0	−.061	−.04	−.02	−.02	−.01
0	.53	.44	.32	.22	.22	0	−.134	−.09	−.04	−.04	−.03
0	−.70	−.58	−.44	−.32	−.30	0	−.286	−.19	−.09	−.08	−.06
0	−.93	−.76	−.61	−.47	−.43	0	−.333	−.22	−.11	−.09	−.07
0	1.12	.92	.78	.62	.55	0	.446	.29	.15	.13	.10
0	1.44	1.18	1.09	.09	.79	0	−.586	−.39	−.20	−.18	−.13
0	1.89	1.55	1.62	1.51	1.22	0	−.664	−.44	−.23	−.21	−.15
2.5	2.74	2.25	2.67	2.65	2.21	0	−1.420	−.93	−.71	−.72	−.47
2.5	3.78	3.11	3.78	3.78	3.26	2.828	2.496	1.64	2.13	2.23	1.55
2.5	3.82	3.14	3.82	3.82	3.30	4.000	3.728	2.45	3.71	3.72	3.09
S.S.E.[a]		6.68	8.52	7.29	4.08			5.20	1.22	1.07	2.76

[a] Sum of squared errors.

Source: θ_0 and x are taken from Dempster [9], Tables 3 and 4.

increase of error for the spherical case where $M=L$.

To get a feeling of how these estimators perform in real applications they have been applied to the examples given by Dempster ([9], pp. 72–73). The results are shown in Table 2.

6. Concluding remarks

The results of numerical experiments given in the preceding section show that our definition of the predictive likelihood of a model is useful. The application seems particularly simple when we start with the models specified by the maximum likelihood estimates of the parameters and define their likelihoods by $\exp(-\mathrm{AIC}/2)$'s. Although we have discussed only the problem of estimation of the mean of a multivariate normal distribution, this problem may be viewed as an approximation to the typical situation of the maximum likelihood estimation of many parameters, and the results reported in this paper suggest wide applicability of the procedure in such situations. The application of the procedure described in this paper to the polynomial regression of (1.1) is particularly straightforward. The application to the autoregressive model fitting is already discussed in Akaike [3] and some computer programs are given in Akaike et al. [4].

Discussing the difficulty of choosing the prior distributions $p_k({}_k\theta)$, Chow [7] noticed the lack of recognition of the distinction between a pre-sample model and a post-sample model among conventional Bayesians. It seems that this distinction corresponds to our distinction of Procedure I (exact Bayes procedure) and Procedure II (quasi-Bayes procedure) in Section 4. The fact that Procedure II could produce practically meaningful results suggests that the possibility of more flexible use of information supplied by the likelihood functions $p_k(x\,|\,{}_k\theta)$ than by the rigid Bayesian approach should not be neglected.

It may be argued that anything accomplished by the definition of the likelihood $L_k(x)$ may be attained by a proper choice of $\pi(k)$. That this is not the case is obvious from the fact that $L_k(x)$ can be defined even when $\pi(k)$ is not specified, as in the cases of θ_{IC}, θ_{UF}, and θ_{UA}. The $\pi(k)$'s used to define θ_{UF} and θ_{UA} are designed to extract useful information supplied by the $L_k(x)$'s. The numerical results of Section 5 show that, at least partially, our definition of $L_k(x)$ provides an answer to the problem discussed by Atkinson [5], the comparison of models specified by some improper prior distributions of the parameters.

One might suspect that the above problem can be treated as the problem of approximation of the exact Bayes procedure, as is done in Davis [8]. By this approach, the heights of the vague prior distributions of the parameters are chosen to produce good approximations to the posterior densities and probabilities. Obviously this is the problem of numerical approximation of Procedure I which cannot be implemented without assuming some proper prior distributions of the parameters. Thus, to apply the results of Davis, we must first settle the practically difficult problem of specifying, at least partially, the proper prior distributions of the parameters. That this difficulty is particularly significant when the models have different numbers of parameters is already demonstrated by the discussion in Section 4.

Although the application was limited to Gaussian model, the quasi-Bayes procedure realized with the use of the predictive likelihood of a data dependent model suggests a natural use of some improper prior distributions in the multi-model situation. It seems that the procedure can lead to a useful practical method of statistical inference. The problem of further characterizing the structure of a model for which the predictive likelihood can be defined remains an interesting open question.

Acknowledgements

The author is grateful to Ms. E. Arahata for programming the necessary computations of the numerical results reported in the paper. This research was partly supported by a grant from the Ministry of Education, Science and Culture.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Akaike, H. (1977). On entropy maximization principle, *Applications of Statistics* (ed. P. R. Krishnaiah), North-Holland, Amsterdam, 27-41.

[2] Akaike, H. (1978). A new look at the Bayes procedure, *Biometrika*, **65**, 53-59.

[3] Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure for autoregressive model fitting, *Biometrika*, **66**, 237-242.

[4] Akaike, H., Kitagawa, G., Arahata, E. and Tada, F. (1979). TIMSAC-78, *Computer Science Monographs*, No. 11, The Institute of Statistical Mathematics, Tokyo.

[5] Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model, *Biometrika*, **65**, 39-48.

[6] Atkinson, A. C. and Cox, David R. (1974). Planning experiments for discriminating between models, (with discussion), *J. Roy. Statist. Soc.*, B, **36**, 321-348.

[7] Chow, Gregory C. (1979). A reconcilation of the information and posterior probability criteria for model selection, *Research Memorandum* No. 234, Econometric Research Program, Princeton University, revised, February, 1979.

[8] Davis, W. W. (1979). Approximate Bayesian predictive distribution and model selection, *J. Amer. Statist. Ass.*, **74**, 312-317.

[9] Dempster, Arther P. (1971). Model searching and estimation in the logic of inference, *Foundation of Statistical Inference* (eds. V. P. Godambe and D. A. Sprott), Holt, Rinehart and Winston of Canada, Toronto, 56-76.

[10] Halpern, Elkan F. (1973). Polynomial regression from a Bayesian approach, *J. Amer. Statist. Ass.*, **68**, 137-143.