Ann. Inst. Statist. Math.
29 (1977), Part A, 9-20

AN OBJECTIVE USE OF BAYESIAN MODELS

HIROTUGU AKAIKE

(Received Mar. 29, 1976)

Introduction

Since the posthumous publication of the work of Bayes [6] on the use of the so-called Bayes' theorem to produce an estimate of the probability of a future happening of an event based on the past observations there has been almost unending discussion on the use of the concept of a prior distribution in statistical inference. Certainly not much discussion is necessary when a prior distribution of the parameters of an objectively defined statistical model has a clearly defined objective meaning. It is only the validity of the use of a subjectively defined prior distribution in connection with the Bayes' theorem that has been questioned. A Bayesian assumes the existence of a prior distribution of the unknown parameters within a model even under the situation where it is clearly understood that the unknown parameters are fixed constants and not considered as realizations of random variables. Although the naturalness of Bayesian solutions of some of the problems of statistical inference can not be questioned, the difficulty of the reconciliation between Bayesians and non-Bayesians remains mainly because of the non-existence of a convincingly objective description of how to choose a prior distribution.

The purpose of this paper is to give an entirely new interpretation of the role played by the prior distribution in a Bayesian model and discuss its implications with the statistical inference. We assume a probability distribution, called a modifier, over the space of parameters within a probabilistic model. The modifier is introduced only for the purpose of constructing a good estimate of the probability distribution of future observations and is considered to be modifying the ordinary likelihood function for this purpose. The utility of a modifier is measured by a criterion of fit of the predictive distribution, an estimate of the probability distribution of future observations, which is obtained by formally assuming the parameters within the model to be random variables distributed according to the posterior distribution obtained by the Bayes' theorem with the modifier as the prior distribution. The criterion of fit is the expected entropy, or the minus Kullback information measure, of the true distribution with respect to the predictive distribution. The criterion clearly shows how the goodness of fit of a predictive distribution depends on the choice of the modifier. It is shown that by using a parametric family of modifiers the method of maximum likelihood can be extended to provide a useful estimate of an optimum modifier.

As a guide to the proper choice of a parametric family of modifiers the situation where the prior distribution is explicitly given is considered. It is shown that the modifier which is identical to the given prior distribution maximizes the mean entropy averaged over the prior distribution. The given prior distribution may either be objective or subjective. At least formally our modifier is an entity which is defined independently of the assumption of existence or non-existence of a prior distribution. The fact that it identifies itself automatically with the prior distribution, when the latter is explicitly given, gives a possible explanation of the nature of the historical confusion between objective and subjective Bayesians.

The implications of the results obtained in the present paper with Robbins' compound decision and empirical Bayes procedures and with the Stein estimator are discussed and the necessity of developing useful parametric families of modifiers is suggested.

Although the problem discussed in this paper serves only as a prototype it clearly demonstrates the necessity and feasibility of an objective use of Bayesian models with modifiers playing the role of data adaptive prior distributions. The basic idea of this data adaptive prior distribution and its evaluation through the expected entropy of the corresponding predictive distribution was first developed in [3]. Somewhat similar idea has been presented by Aitchison [1], but this is not concerned with the adjustment of prior distributions by data.

1. Predictive distribution and entropy

A very general problem of statistical inference is concerned with the prediction of the probability distribution of some future observations. In this paper we assume that the object of statistical inference is to provide an estimate of the distribution of a future observation of a vector random variable $\boldsymbol{y}$ from the observation of $\boldsymbol{x}$ where both $\boldsymbol{x}$ and $\boldsymbol{y}$ are distributed according to one and the same distribution. It is assumed that the distribution of $\boldsymbol{y}$ has a density $g(\boldsymbol{y})$ with respect to a measure $d\boldsymbol{y}$. It is also assumed that $g(\boldsymbol{y})$ is a member of a parametric family $\{f(\boldsymbol{y}|\boldsymbol{\theta});\ \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$, i.e., there exists a $\boldsymbol{\theta}_0$ in $\boldsymbol{\Theta}$ such that $g(\boldsymbol{y})=f(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Further it is assumed that the true parameter $\boldsymbol{\theta}_0$ is unknown and the problem is to find a density $h(\boldsymbol{y}|\boldsymbol{x})$ of $\boldsymbol{y}$ as a function of $\boldsymbol{x}$ which will be a good estimate of $g(\boldsymbol{y})$. Here $h(\boldsymbol{y}|\boldsymbol{x})$ is not necessarily a member of the family $\{f(\boldsymbol{y}|\boldsymbol{\theta});\ \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$.

As the criterion of goodness of fit of a probability density function $h(\boldsymbol{y})$ as an approximation to $g(\boldsymbol{y})$ we use the entropy of $g(\boldsymbol{y})$ with respect to $h(\boldsymbol{y})$ which is defined by

$$B(g;h)=-\int\log\left\{\frac{g(\boldsymbol{y})}{h(\boldsymbol{y})}\right\}g(\boldsymbol{y})d\boldsymbol{y}\ .$$

Hereafter it will tacitly be assumed that all the integrals appearing in the discussion have finite values. The neg-entropy $-B(g;h)$ is identical to the Kullback information measure $I(g;h)$ [9]. The greater the entropy $B(g;h)$, the degree of approximation of $h(\boldsymbol{y})$ to $g(\boldsymbol{y})$ is considered to be higher. This is a natural extension of Boltzmann's interpretation of the thermodynamical concept of entropy as the logarithm of the probability of obtaining a statistical distribution and a mathematical justification of the present interpretation can be found in Sanov [14]. $B(g;h)$ is non-positive. The criterion of fit of an estimate $h(\boldsymbol{y}|\boldsymbol{x})$ to the distribution $g(\boldsymbol{y})$ is given by the expected entropy $E_xB(g;h(\ |\boldsymbol{x}))$, where E_x denotes expectation with respect to the distribution of $\boldsymbol{x}$.

Bayesians assume the existence of a probability distribution of $\boldsymbol{\theta}$ which is here specified by a density function $p(\boldsymbol{\theta})$. When x is observed the prior density $p(\boldsymbol{\theta})$ is transformed into a density $p(\boldsymbol{\theta}|\boldsymbol{x})$ by the rule of Bayes

$$p(\boldsymbol{\theta}|\boldsymbol{x})=\frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}\ .$$

By using the posterior density $p(\boldsymbol{\theta}|\boldsymbol{x})$ an estimate of $g(\boldsymbol{y})=f(\boldsymbol{y}|\boldsymbol{\theta}_0)$ is defined by

$$h(\boldsymbol{y}|\boldsymbol{x})=\int f(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}\ .$$

Following Roberts [11] we will call $h(\boldsymbol{y}|\boldsymbol{x})$ the predictive distribution of $\boldsymbol{y}$ implied by the posterior distribution. If in the definition of $h(\boldsymbol{y}|\boldsymbol{x})$ $p(\boldsymbol{\theta}|\boldsymbol{x})$ is replaced by $p(\boldsymbol{\theta})$ we get the predictive distribution of $\boldsymbol{y}$ implied by the distribution $p(\boldsymbol{\theta})$. The distribution specified by $f(\boldsymbol{x}|\boldsymbol{\theta})$ is sometimes called the data distribution.

Here we take the attitude that except for its mathematical characteristics as a probability distribution the prior $p(\boldsymbol{\theta})$ need not have any meaning as an objective or subjective probability distribution and assume that its role is to provide a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ which will give a good predictive distribution $h(\boldsymbol{y}|\boldsymbol{x})$. The goodness of a predictive distribution is evaluated by the expected entropy $E_xB(g; h(\ |\boldsymbol{x}))$. To discriminate the present use of $p(\boldsymbol{\theta})$ from the use as an ordinary objective or subjective prior distribution we call $p(\boldsymbol{\theta})$ a modifier, $f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ the modified likelihood and $\int f(x|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ the aggregate likelihood. Since there is no possibility of conceptual confusion $p(\boldsymbol{\theta}|\boldsymbol{x})$ which is defined as the ratio of a modified likelihood to the aggregate will be called the posterior density.

2. Entropy maximization

In this section we treat a simple but typical example of the use of a Bayesian model. We assume that $\boldsymbol{y}=(y_1, y_2, \cdots, y_N)$ and $\boldsymbol{x}=(x_1, x_2, \cdots, x_N)$ denote independent observations from an N-dimensional Gaussian distribution $N(\boldsymbol{\theta}_0, \boldsymbol{I}_N)$, where $\boldsymbol{\theta}_0=(m_1, m_2, \cdots, m_N)$ and $\boldsymbol{I}_N$ denotes an $N\times N$ identity matrix. The family of the data distribution is $\{N(\boldsymbol{\theta}, \boldsymbol{I}_N); \boldsymbol{\theta} \in R^N\}$, where R^N denotes a real N-dimensional vector space, and the problem is to find a useful predictive distribution of $\boldsymbol{y}$ by using the observation $\boldsymbol{x}$. The value of the true parameter $\boldsymbol{\theta}_0$ is fixed but unknown.

Although we know that we are concerned here with a situation where the parameter is fixed at $\boldsymbol{\theta}_0$ we still somewhat arbitrarily assume a distribution $p(\boldsymbol{\theta})$, a modifier, which is an N-dimensional Gaussian distribution $N(0, \sigma^2\boldsymbol{I}_N)$ and try to find out whether the formal application of Bayes' theorem can produce a useful posterior distribution. It can easily be seen that in the present case the posterior distribution is an N-dimensional Gaussian distribution $N(c\boldsymbol{x}, c\boldsymbol{I}_N)$ with $c=\sigma^2/(1+\sigma^2)$. The predictive distribution of $\boldsymbol{y}$ implied by this posterior distribution is given by $N(c\boldsymbol{x}, (1+c)\boldsymbol{I}_N)$ and we have

$$\log h(\boldsymbol{y}|\boldsymbol{x})=-\frac{1}{2}\sum_{i=1}^{N}\left\{\log 2\pi+\log(1+c)+\frac{(y_i-cx_i)^2}{1+c}\right\},$$

and

$$E_y \log h(\boldsymbol{y}|\boldsymbol{x})=-\frac{1}{2}\sum_{i=1}^{N}\left\{\log 2\pi+\log(1+c)+\frac{1+(m_i-cx_i)^2}{1+c}\right\},$$

where E_y denotes expectation with respect to the distribution $f(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Since we have

$$E_y \log f(\boldsymbol{y}|\boldsymbol{\theta}_0)=-\frac{1}{2}\sum_{i=1}^{N}\{\log 2\pi+1\}$$

the expected entropy is given by

$$\begin{aligned}&E_xB(f(\ |\boldsymbol{\theta}_0); h(\ |\boldsymbol{x}))\\ &\quad=E_xE_y(\log h(\boldsymbol{y}|\boldsymbol{x})-\log f(\boldsymbol{y}|\boldsymbol{\theta}_0))\end{aligned}$$

$$=-\frac{N}{2}\left\{\log(1+c)+\frac{1}{1+c}-1+\frac{c^2+(1-c)^2\overline{m^2}}{1+c}\right\},$$

where $\overline{m^2}=(1/N)\sum m_i^2$. Here we consider a family $\{p(\boldsymbol{\theta}|\sigma^2):\ 0\leqq\sigma^2<\infty\}$ of modifiers, where $p(\boldsymbol{\theta}|\sigma^2)$ is defined by $N(0,\sigma^2\boldsymbol{I}_N)$. The expected entropy of $g(\boldsymbol{y})=f(\boldsymbol{y}|\boldsymbol{\theta}_0)$ with respect to the predictive distribution $h(\boldsymbol{y}|\boldsymbol{x})$ is maximized at $c=c_{\text{opt}}=\overline{m^2}/(1+\overline{m^2})$ or at $\sigma^2=\sigma^2_{\text{opt}}=\overline{m^2}$. This result shows that for the present family of modifiers the one with the value of σ^2 equal to $\overline{m^2}$ produces the best result. If the Bayesian procedure is applied with priors $N(0,\sigma^2\boldsymbol{I}_N)$ by many independent individuals to the present problem with a fixed value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$, the group of those who use the prior distribution with σ^2 close to $\overline{m^2}$ will be more successful in estimating $f(\boldsymbol{y}|\boldsymbol{\theta}_0)$. Thus even a Bayesian will be interested in exploring the possibility of profitably adjusting the prior distribution, using the information provided by the data $\boldsymbol{x}$.

A conservative Bayesian may not like the idea of adjusting the prior distribution by data. In this case he must use a fixed σ^2 or c. It is assumed here that the use of $p(\boldsymbol{\theta}|\sigma^2)$ is sanctioned. It can then be seen that for any value of c different from 1 the corresponding predictive distribution produces a poor estimate, i.e., $E_xB(f(\ |\boldsymbol{\theta}_0);h(\ |\boldsymbol{x}))$ tends to be significantly negative, when $\overline{m^2}$ is very large. The modifier with $c=1$ provides a minimax solution to this case. The solution is given by the limiting case $\sigma^2=\infty$, which corresponds to the non-informative prior distribution. The value of the expected entropy $E_xB(f(\ |\boldsymbol{\theta}_0);$ $h(\ |\boldsymbol{x}))$ for $c=1$ is

$$-\frac{N}{2}\log 2,$$

which should be compared with the value for the optimal choice $c=\overline{m^2}$

$$-\frac{N}{2}\log\left\{2-\frac{1}{1+\overline{m^2}}\right\}.$$

If $\overline{m^2}$ is large compared with 1, the value of the variance of x_i, the difference between the two choices is small, but otherwise not. From the present point of view of the use of a modifier, the uncritical recommendation of the non-informative prior distribution in a Bayesian approach is unjustified.

The expected entropy of the predictive distribution $f(\boldsymbol{y}|\boldsymbol{x})$ defined by putting $\boldsymbol{\theta}$ of $f(\boldsymbol{y}|\boldsymbol{\theta})$ equal to $\boldsymbol{x}$, which is the maximum likelihood estimate of $\boldsymbol{\theta}_0$, is

$$E_xB(f(\ |\boldsymbol{\theta}_0);f(\ |\boldsymbol{x}))=-\frac{N}{2}.$$

Comparing this with the former results it is clear that even the Bayesian procedure based on the non-informative prior distribution is producing a better predictive distribution than $f(\boldsymbol{y}|\boldsymbol{x})$ which is the predictive distribution implied by the maximum likelihood estimate of $\boldsymbol{\theta}_0$. The routine approximation of a posterior distribution by a distribution concentrated at its mode produces a predictive distribution $f(\boldsymbol{y}|c\boldsymbol{x})$, for which the expected entropy is evaluated as

$$E_x B(f(\ |\boldsymbol{\theta}_0); f(\ |c\boldsymbol{x})) = -\frac{N}{2}\{c^2+(1-c)^2\overline{m^2}\} .$$

In our terminology the mode of the posterior distribution defines the maximum modified likelihood estimate of $\boldsymbol{\theta}_0$. The optimal choice of c is again given by $c=c_{\text{opt}}=\overline{m^2}/(1+\overline{m^2})$ and we have

$$E_x B(f(\ |\boldsymbol{\theta}_0); f(\ |c_{\text{opt}}\boldsymbol{x})) = -\frac{N}{2}c_{\text{opt}} .$$

The minimax solution to this case is given by $f(\boldsymbol{y}|\boldsymbol{x})$, the predictive distribution implied by the maximum likelihood estimate of $\boldsymbol{\theta}_0$.

3. Extended use of maximum likelihood estimates

Here we explore the feasibility of using the data $\boldsymbol{x}$ to choose a modifier which approximates the optimum choice within a parametric family. It will be shown that this will at least asymptotically be realized by using the parameters which maximize the aggregate likelihood.

Generally, for a parametric modifier $p(\boldsymbol{\theta}|\boldsymbol{\delta})$ with a vector of parameters $\boldsymbol{\delta}$, the predictive distribution of $\boldsymbol{x}$ implied by the modifier is given by

$$k(\boldsymbol{x}|\boldsymbol{\delta}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\delta})d\boldsymbol{\theta} .$$

When $\boldsymbol{x}$ is observed $k(\boldsymbol{x}|\boldsymbol{\delta})$ defines the aggregate likelihood of the Bayesian model defined by the data distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$ and the modifier $p(\boldsymbol{\theta}|\boldsymbol{\delta})$. It should be remembered that in the present definition of the aggregate likelihood the modifier $p(\boldsymbol{\theta}|\boldsymbol{\delta})$ need not represent any objective probability distribution.

For the case treated in the preceding section we have

$$\log k(\boldsymbol{x}|\sigma^2) = -\frac{N}{2}\left(\log 2\pi + \log(1+\sigma^2) + \frac{\overline{x^2}}{1+\sigma^2}\right) ,$$

where $\overline{x^2}=(1/N)\sum x_i^2$. The value of σ^2 which maximizes the aggregate likelihood is given by

$$\hat{\sigma}^2=\begin{cases}\overline{x^2}-1 & \text{if } \overline{x^2}>1\,, \\ 0 & \text{otherwise}\,.\end{cases}$$

In a situation where N grows indefinitely and $\overline{m^2}$ tends to a constant m_0^2, $\overline{x^2}$ will converge to m_0^2+1 with probability one, where $\overline{m^2}$ is the average of the squared mean values. This shows the potential of the maximum aggregate likelihood estimate $\hat{\sigma}^2$ as an estimate of $\sigma_{\text{opt}}^2=\overline{m^2}$. It must be remembered that the present modifier $p(\boldsymbol{\theta}\,|\,\sigma^2)$ was chosen quite arbitrarily and σ^2 was not supposed to have any objectively defined "true value". It was only through the process of maximizing the expected entropy that the optimal value $\sigma_{\text{opt}}^2=\overline{m^2}$ was introduced. It is interesting and also quite natural that the method of maximum likelihood applied to the aggregate likelihood automatically leads to a meaningful estimate of σ_{opt}^2. The gist of the present procedure lies in the reduction of the number of free parameters within the model, attained by switching from the original likelihood to the aggregate likelihood.

The above procedure of maximum aggregate likelihood corresponds to the method of type II maximum likelihood of Good [7]. Good considers the procedure a type of Bayes/non-Bayes compromise and does not consider objective justification of its validity.

4. Objective and subjective priors

Although it has been shown that the method of maximum aggregate likelihood can produce a useful estimate of the optimum modifier $p(\boldsymbol{\theta}\,|\,\sigma_{\text{opt}}^2)$ the outstanding problem is the choice of the functional form of the modifier. When we consider the situation where the true parameter $\boldsymbol{\theta}_0$ is a fixed constant and the condition of estimability is ignored the problem of optimal functional form of the modifier has only a trivial answer, the distribution concentrated at $\boldsymbol{\theta}_0$. When $\boldsymbol{\theta}_0$ is distributed according to an objectively defined prior $p(\boldsymbol{\theta}_0)$ the average excepted entropy of a predictive distribution $h(\boldsymbol{y}\,|\,\boldsymbol{x})$ is defined by

$$B(f\,;h|\,p)=E_0E_xB(f(\ \,|\boldsymbol{\theta}_0)\,;h(\ \,|\boldsymbol{x}))\,,$$

where E_0 denotes expectation with respect to the prior $p(\boldsymbol{\theta}_0)$. It is easy to see, and in fact is shown by Aitchison [1], that the predictive distribution implied by the posterior $p(\boldsymbol{\theta}_0\,|\,\boldsymbol{x})$, obtained from the prior $p(\boldsymbol{\theta}_0)$, minimizes $B(f\,;h\,|\,p)$. This fact is a direct consequece of the following relation:

$$B(f\,;h\,|\,p)=-\int\!\!\int\!\!\int\log\left\{\frac{f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_0)}{h(\boldsymbol{y}\,|\,\boldsymbol{x})}\right\}f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_0)f(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)d\boldsymbol{y}d\boldsymbol{x}d\boldsymbol{\theta}_0$$

$$=-\int\int \log\{f(\boldsymbol{y}|\boldsymbol{\theta}_0)\}f(\boldsymbol{y}|\boldsymbol{\theta}_0)d\boldsymbol{y}p(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0$$
$$+\int\left[\int \log\{h(\boldsymbol{y}|\boldsymbol{x})\}\left\{\int f(\boldsymbol{y}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0|\boldsymbol{x})d\boldsymbol{\theta}_0\right\}d\boldsymbol{y}\right]p(\boldsymbol{x})d\boldsymbol{x}\ ,$$

where $p(\boldsymbol{\theta}_0|\boldsymbol{x})=f(\boldsymbol{x}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)/p(\boldsymbol{x})$, $p(\boldsymbol{x})=\int f(\boldsymbol{x}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0$ and it is assumed that $h(\boldsymbol{y}|\boldsymbol{x})$ is independent of $\boldsymbol{\theta}_0$. The last term attains its maximum with $h(\boldsymbol{y}|\boldsymbol{x})=\int f(\boldsymbol{y}|\boldsymbol{\theta}_0)p(\boldsymbol{\theta}_0|\boldsymbol{x})d\boldsymbol{\theta}_0$. This result shows that *if we are interested in the average goodness of fit of our predictive distribution we should use the prior $p(\boldsymbol{\theta})$ as our modifier.*

Here we must also note that if in the above discussion $p(\boldsymbol{\theta}_0)$ had been replaced by a subjective prior distribution, or a weight distribution introduced for the purpose of assessment of predictive distributions, we would have arrived at one and the same conclusion.

The above observations clarify part of the confusion between the objective and subjective Bayesians. *The fact that both objective and subjective prior distributions can play the role of an optimum modifier could have been quite effective in obscuring the distinction between the objective and subjective prior distributions.* Rather than the distinction between the objective and subjective prior distributions the distinction between the cases with known and unknown prior distributions is more essential. Our discussion of modifier is more concerned with the case where the prior distribution is not explicitly given than with the simple and uninteresting case where the prior distribution is known.

Though uninteresting in itself the analysis of the cases where the prior distributions are assumed to be known gives us some idea as to how the functional form of the modifier should be chosen. The lesson from the present analysis of the cases with known prior distributions is that we should choose a functional form of the modifier which either would not seriously contradicts our knowledge of the objective prior distribution, or would not be too much awkward as the weight distribution for the assessment of the predictive distribution.

5. Discussions

Here we will discuss some of the implications of the results obtained in the preceding sections.

a. *Complete freedom in the choice of statistical models*

It has already been shown that *a subjectively contrived modifier can produce an objectively meaningful predictive distribution.* This results clearly demonstrates the necessity and use of subjective element in statistical inference. In practical situations even the exact form of the data distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$ will rarely be known precisely. Every statistical model used in a inference situation must be considered as a subjectively chosen approximation to the true structure. The explicit recognition of the role of a modifier as a device for the production of a useful predictive distribution completely frees us from the historically limited views of the prior distribution. Even if we are not Bayesians we can now freely use a Bayesian model. It is whether the assumed Bayesian model can produce an objectively useful result or not that matters.

b. *Relation to Robbins' and Stein's works*

In the discussion of the compound decision problem Robbins [12] carefully discriminates the situation of a compound decision problem from the situation where a prior distribution exists. The latter is treated by the empirical Bayes procedure [13]. The distinction has also been stressed by Neyman [10]. The discussion in the present paper suggests that the distinction would be immaterial. The analyses in the preceding sections give us an idea of *what would happen when a properly formulated empirical Bayes procedure is applied to a situation where a prior distribution does not exist.*

Stein [15] discussed the problem of inadmissibility of the maximum likelihood estimates of the means of an independently and identically distributed Gaussian random variables with respect to the sum of squared error and suggested the use of shrinked estimates. Taking into account the relation $B(f(\ |\boldsymbol{\theta}_0); f(\ |c\boldsymbol{x}))=-(1/2)\{\sum (m_i-cx_i)^2\}$, it is obvious that the problem is directly connected with our discussion of the predictive distribution $f(\boldsymbol{y}|c\boldsymbol{x})$ implied by the mode of the posterior distribution. Since the expected entropy of $f(\boldsymbol{y}|c\boldsymbol{x})$ with the optimal choice of $c=c_{\text{opt}}=\overline{m^2}/(1+\overline{m^2})$ is given by $-(N/2)c_{\text{opt}}$, only those cases where $\overline{m^2}$ is comparable to or less than 1, the value of the variance of observations, will deserve to a careful analysis. By replacing $\overline{m^2}$ by the maximum aggregate likelihood estimate $\hat{\sigma}^2$ given in Section 3 we get a naturally non-negative shrinkage factor defined by $c=1-(1/\overline{x^2})$ for $\overline{x^2}>1$, 0 otherwise. It can be seen from the result of James and Stein ([8], p. 365) that the estimate $c\boldsymbol{x}$ defined by this factor is better than the original maximum likelihood estimate $\boldsymbol{x}$ at least for $N\geqq 4$, i.e., $E_x \sum (m_i-cx_i)^2 \leqq N$ for $N\geqq 4$.

If in the discussion of Section 2 the posterior distribution $N(c\boldsymbol{x}, c\boldsymbol{I}_N)$ is replaced by $N(c\boldsymbol{x}, \boldsymbol{I}_N)$ we get a predictive distribution $h(\boldsymbol{y}|c\boldsymbol{x})$ for which we have

$$\log h(\boldsymbol{y}|c\boldsymbol{x})=-\frac{1}{2}\sum_{i=1}^{N}\left\{\log 2\pi+\log 2+\frac{(y_i-cx_i)^2}{2}\right\} .$$

The expected entropy of this predictive distribution is given by

$$E_xB(f(\ |\boldsymbol{\theta}_0);h(\ |c\boldsymbol{x}))=-\frac{N}{2}\left\{\log 2-\frac{1}{2N}\left(N-E_x\sum_{i=1}^{N}(m_i-cx_i)^2\right)\right\}.$$

From the above stated result of James and Stein we have

$$E_xB(f(\ |\boldsymbol{\theta}_0);h(\ |c\boldsymbol{x}))\geqq-\frac{N}{2}\log 2.$$

As was shown in Section 2, the right-hand side is equal to the expected entropy of the predictive distribution implied by the posterior distribution obtained from an non-informative uniform prior distribution. Thus for $N\geqq 4$ the present $h(\boldsymbol{y}|c\boldsymbol{x})$ is clearly better than the predictive distribution given by the uniform prior distribution.

A discussion of the Stein's estimator from the standpoint of entropy maximization without using Bayesian approach is presented in [4].

c. *Comments on the practical use of the model treated in Sections* 2 *and* 3

In interpreting the results of Sections 2 and 3 in relation to practical applications, we can consider the situation where $\boldsymbol{\theta}_0$ is the vector of parameters within a general statistical model and $\boldsymbol{x}$ is the maximum likelihood estimate of $\boldsymbol{\theta}_0$. This situation can be realized at least approximately by properly transforming the original model so that the Fisher information matrix becomes an identity matrix. In this case the number N of the components of $\boldsymbol{\theta}$ is fixed and does not justify the application of the asymptotic discussion of the empirical Bayes procedure. There is another parameter M here, the size of the original sample which produced the maximum likelihood estimate $\boldsymbol{x}$. The value $\overline{m^2}$ represents the signal to noise ratio and tends to be higher as M is increased.

It will often be the case that some of the m_i's are with magnitudes significantly higher than 1 and some of the rest are with magnitudes significantly lower than 1. It would seem inadequate to treat these situations with the modifier used in Section 2. The necessary modification of the modifier to treat this type of situation is rather straightforward and is of great practical importance. A modifier to treat this type of situations is now being empirically tested and is producing encouraging results. This will be discussed elsewhere [5]. The analysis of the relation between this type of approach and the minimum AIC estimation procedure [2] would be a subject of further study.

6. Conclusion

It has been demonstrated by a concrete example that an objective use of a Bayesian model with a modifier replacing the role of the prior distribution can be quite useful. The most important observation is the explicit recognition of the use of a modifier which can be applied irrespectively of the existence or non-existence of the prior distribution. The choice of the functional form of a modifier could depend on the available prior information and the convenience of its use but the parameters are adjusted by the data through an objectively defined procedure. By this approach a wide reconciliation between the Bayesian and non-Bayesian approaches is made possible. The development of useful parametric families of modifiers and the analysis of the statistical characteristics of the extended maximum likelihood estimates would be the most important and fruitful subjects of future study.

As a final comment it must be mentioned here that the maximum aggregate likelihood estimate of σ^2 can be viewed as the mode of the posterior distribution of σ^2 for the improper uniform prior distribution. This observation and the results of the preceding sections thus confirm the objective utility of a fully Bayesian approach. The only problem with the Bayesian approach was how to prove its objective utility. The point of view of maximization of the expected entropy of the predictive distribution prepared a general principle for the evaluation and development of this type of models.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Aitchison, J. (1975). Goodness of prediction fit, *Biometrika*, **62**, 547-554.

[2] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, **AC-19**, 716-723.

[3] Akaike, H. (1975). Concept of information quantity and the theory of statistics, *Preprints of Fall Meeting of Mathematical Society of Japan*, October 1975, 73-83 (in Japanese).

[4] Akaike, H. (1975). An extension of the method of maximum likelihood and the Stein's problem, *Research Memo*. No. 84, Inst. Statist. Math.

[5] Akaike, H. (1976). On entropy maximization principle, A paper presented at the Symposium on Applications of Statistics, Dayton, Ohio, June 14-18, 1976.

[6] Bayes, T. (1763). An essay towards solving a problem in the doctorine of chances, *Phil. Trans.*, **53**, 370-418. Reprinted in Thomas Bayes's essay toward solving a problem in the doctorine of chances, *Biometrika*, **45** (1958), 293-315.

[7] Good, I. J. (1965). *The Estimation of Probabilities*, M.I.T. Press, Cambridge.

[8] James, W. and Stein, C. M. (1961). Estimation with quadratic loss function, *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, **1**, 361-379.

[9] Kullback, S. (1959). *Information Theory and Statistics*, John Wiley, New York.

[10] Neyman, J. (1962). Two breakthroughs in the theory of statistical decision making, *Rev. Int. Statist. Inst.*, **30**, 11-27.

[11] Roberts, H. V. (1965). Probabilistic prediction, *J. Amer. Statist. Ass.*, **60**, 50-62.

[12] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems, *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, 131-148.

[13] Robbins, H. (1956). An empirical Bayes approach to statistics, *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, 1, 157-164.

[14] Sanov, I. N. (1961). On the probability of large deviations of random variables, *IMS and AMS Selected Translation in Mathematical Statistics and Probability*, 1, 213-244.

[15] Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, 1, 197-206.