Ann. Inst. Statist. Math.
Vol. 40, No. 3, 587-602 (1988)

A BAYESIAN APPROACH TO NONPARAMETRIC TEST PROBLEMS*

YOSIYUKI SAKAMOTO AND MAKIO ISHIGURO

The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku, Tokyo 106, Japan

(Received August 13, 1986; revised September 22, 1987)

Abstract. We propose an alternative approach to the classical "nonparametric" test problems, such as the goodness of fit test and the two-sample "nonparametric" test. In this approach, those problems are reviewed from the viewpoint of the estimation of the underlying population distributions and are formulated as the problem of model selection between Bayesian models which were recently proposed by the present authors. The model selection can be easily realized by choosing a model with the smallest ABIC, Akaike Bayesian information criterion. The approach provides the estimates of the density of the underlying population distribution(s) of any shape as well as the evaluation of the goodness of fit or the check of homogeneity of distributions. The practical utility of the present procedure is demonstrated by numerical examples. The difference in behavior between the present procedure and a density estimator GALTHY proposed by Akaike and Arahata is also briefly discussed.

Key words and phrases: Goodness of fit test, two-sample nonparametric test, Bayesian model, smoothing prior, nonparametric density estimator, model selection, ABIC, AIC, multinomial logistic transformation, B-spline.

1. Introduction

We must frequently carry out statistical analyses under conditions that we know very little about the shape of the population distribution from which the samples are drawn. Many two-sample "nonparametric" test procedures have been proposed and claimed to be valid for samples from continuous population distributions of any shape. Nevertheless, most of conventional test procedures are based on rank statistic, and they cannot provide any information on the shape of the population distributions, even

*This paper was originally read at the Conference on "Graphical Models to Analyze Structures" (Organizer: N. Wermuth, Johannes Gutenberg University), June 30-July 2, 1986, Wiesbaden, West Germany.

when much data are accumulated. The rank statistic prevents the data analyst from making the most of the information supplied by the data.

Moreover, in the case of the chi-square goodness of fit test, the conventional test procedure has a serious problem in that there is no reasonable procedure which gives an initial classification for a set of continuous observations.

Following Akaike (1977), the principle of entropy maximization is stated as follows: formulate the object of statistical inference as the estimation of a probability distribution from a set of observations and attempt to determine the probability distribution that will maximize the expected entropy. To put this idea into practice, we must develop an estimation procedure for a population distribution of any shape. Akaike (1977) and Akaike and Arahata (1978) have developed the program GALTHY for an automated density estimation and have shown the feasibility of a parametric approach to the goodness of fit test problem. However, GALTHY has two problems requiring improvement. One is that it sometimes produces final estimates with spurious peaks due to the parameterization. It is common to ordinary parametric approaches that the estimate obtained is very much dependent on the appropriateness of the assumed model class. The other is that the final estimate produced by GALTHY is obtained by mixing estimates given by several parametric models and no estimate of measure for the goodness of fit of the final model itself is provided. This causes, for example, GALTHY to deal imperfectly with the "nonparametric" test problem. These two problems can be solved by using a Bayesian approach proposed recently by Ishiguro and Sakamoto (1984).

The purpose of the present paper is to rectify these limitations of GALTHY, and to show the feasibility of an alternative approach to the goodness of fit test problem and to the two-sample "nonparametric" test problem from the point of view of the construction of Bayesian models and their evaluation by the Akaike Bayesian information criterion, ABIC (Akaike (1980)).

For the convenience of the reader, we shall briefly review the Bayesian density estimator in the following section.

2. A Bayesian density estimator

2.1 *A density estimator GALTHY*

Our Bayesian approach borrows an idea from GALTHY: the adoption of the transformation $y = Q(x)$ of original data x, where Q is a properly chosen distribution function (see also Neyman (1937)). This transformation converts the sample space of the original data x into the closed interval $[0, 1]$. If we use $r^*(y)$ to denote the true density function of the data y on the interval $[0, 1]$, then the density function of x is given by

$$f(x) = r^*(Q(x))q(x)\,, \tag{2.1}$$

where $q(x)$ is the density function corresponding to the distribution function $Q(x)$.

In GALTHY, to estimate $r^*(y)$, the fitting of a parametric model defined by

$$r_K(y) = \exp\left\{\sum_{k=0}^{K} a_k y^k\right\} \qquad K = 0, 1, \ldots\,, \tag{2.2}$$

is first tried using the method of maximum likelihood for each K. The final estimate of the density function $r^*(y)$ is found by the weighted average of the estimates of those models with a weight proportional to exp $\{(-1/2)\text{AIC}\}$ for each model.

2.2 *A Bayesian model*

For simplicity, we assume that the density is defined on the closed interval $[0, 1]$. The handling of the general case where the support of the density function is not necessarily bounded is discussed in the next section.

The basic assumption here is that $r^*(y)$ is well approximated by a piecewise constant function defined by

$$r(y) = Cp_j = \frac{C \exp(h_j/C)}{\sum_{k=1}^{C} \exp(h_k/C)} \tag{2.3}$$

$$\text{for} \quad d_{j-1} \le y < d_j \quad j = 1, \ldots, C-1,$$

$$d_{j-1} \le y \le d_j \quad j = C\,,$$

where C is the number of cells and $\{d_j\}$ are defined by

$$d_j = \frac{j}{C} \qquad j = 0, \ldots, C\,, \tag{2.4}$$

and h_j is a parameter which satisfies the relation

$$h_C = -\sum_{j=1}^{C-1} h_j\,. \tag{2.5}$$

Note that the model (2.3) is flexible enough if it is possible to set C very large. Given a set of data $\{y_i;\ 0 \le y_i \le 1, i = 1, \ldots, n\}$, the likelihood of the model (2.3) as a function of $h = (h_1, \ldots, h_{C-1})^T$ is given by

$$L(h) = \prod_{i=1}^{n} r(y_i|h) = \prod_{j=1}^{C} \left\{ \frac{C \exp(h_j/C)}{\sum_{k=1}^{C} \exp(h_k/C)} \right\}^{n_j}, \tag{2.6}$$

where n_j is the number of observations which satisfy $d_{j-1} \leq y_i < d_j$ ($d_{C-1} \leq y_i \leq d_C$ for $j = C$). Although the maximum likelihood estimate of h is easily obtained, it is unstable or noisy when C is large compared with n. To obtain a smooth estimate of h we introduce the prior density of the parameter h defined by

$$\pi(h|v^2, h_{-1}, h_0) = \prod_{j=1}^{C} \frac{1}{\sqrt{2\pi}\, v} \exp\left\{ -\frac{1}{2v^2} (h_j - 2h_{j-1} + h_{j-2})^2 \right\}, \tag{2.7}$$

where v^2, h_{-1} and h_0 are adjustable hyperparameters. When the values of these hyperparameters are fixed, we define the estimate of the parameter h as the mode of the posterior density which is proportional to $L(h) \cdot \pi(h|v^2, h_{-1}, h_0)$.

2.3 *ABIC*

In this approach the selection of those hyperparameters, v^2, h_{-1} and h_0, is crucial. Akaike (1980) proposed the use of the likelihood of a Bayesian model as a criterion for the choice of such hyperparameters. In this case the likelihood of the Bayesian model is defined by

$$\int L(h)\pi(h|v^2, h_{-1}, h_0)dh\,. \tag{2.8}$$

Considering the compatibility with the statistic AIC (Akaike (1973)), Akaike defined the statistic ABIC by

$$\begin{aligned} \text{ABIC} = &-2 \log (\text{maximum likelihood of a Bayesian model}) \\ &+ 2(\text{number of estimated hyperparameters})\,. \end{aligned} \tag{2.9}$$

For the present case, the ABIC is given by

$$\text{ABIC} = -2 \log \int L(h)\pi(h|v^2, h_{-1}, h_0)dh + 2 \times 3\,. \tag{2.10}$$

The values of these three hyperparameters, v^2, h_{-1} and h_0, are to be chosen so that they minimize the ABIC. To find the ABIC value we must carry out the integration of (2.8), which is difficult. We avoid this difficulty by approximating $\log L(h)\pi(h|v^2, h_{-1}, h_0)$ with its Taylor expansion up to the second order term around the maximizing point $\hat{h}$. For further details of this technique, see Ishiguro and Sakamoto (1984).

Note that (2.8) can be viewed as a particular model obtained by

specifying the three hyperparameters for $\{y_i\}$. Thus, the determination of the values of these hyperparameters by the maximization of (2.8) is no more than the method of maximum likelihood with respect to the hyperparameters.

2.4 *Estimation procedure for the general data set*

Given original data $x_1,\ldots,x_n$, each of which does not necessarily belong to the interval [0, 1], we adopt the following estimation procedure.

Step 1: To transform those data $x_1,\ldots,x_n$ by $y = Q(x)$.

Step 2: To find the Bayesian estimate $\hat{r}(y|h)$ of y.

Step 3: To obtain a second-order spline estimate $\tilde{\tilde{r}}(y|h)$ by substituting the basis of second-order B-spline of $\hat{r}(y|h)$ for the corresponding 0-th order B-spline basis. Here, we keep the coefficients intact.

Step 4: To estimate the probability density function of x by

$$\hat{f}(x) = \tilde{\tilde{r}}(Q(x)|h)q(x)\,. \tag{2.11}$$

In the Step 1, $Q(x)$ is chosen by the minimum AIC procedure among familiar distributions, such as the normal distribution, the log-normal distribution and the exponential distribution. We use $\hat{q}(x)$ to denote the density function chosen by this procedure and temporarily call it a reference distribution.

The Step 3 is laid down to avoid the difficulty of the final estimate of $f(x)$ being saw-toothed if $\hat{r}(y|h)$ is adopted instead of $\tilde{\tilde{r}}(y|h)$ in (2.11). Here the second-order spline function $\tilde{\tilde{r}}(y|h)$ is given by

$$\tilde{\tilde{r}}(y|h) = \sum_{j=1}^{C} C\hat{p}_j S_j(y)\,, \tag{2.12}$$

where $\hat{p}_j$ is obtained by substituting $\hat{h}_j$ for h_j in the second equation of (2.3) and $S_j(y)$ is defined by

$$S_j(y) = \begin{cases} 0 & y < b_j - \dfrac{3}{2C}\,, \\ \dfrac{1}{2}C^3\left(y - b_j + \dfrac{3}{2C}\right)^2 & b_j - \dfrac{3}{2C} \le y < b_j - \dfrac{1}{2C}\,, \\ -C^3(y-b_j)^2 + \dfrac{3}{4}C & b_j - \dfrac{1}{2C} \le y < b_j + \dfrac{1}{2C}\,, \\ \dfrac{1}{2}C^3\left(y - b_j - \dfrac{3}{2C}\right)^2 & b_j + \dfrac{1}{2C} \le y < b_j + \dfrac{3}{2C}\,, \\ 0 & y \ge b_j + \dfrac{3}{2C}\,. \end{cases} \tag{2.13}$$

Here b_j denotes the mid-point of the j-th class.

We define the ABIC for the original data $x_1, \ldots, x_n$ by

$$\text{ABIC} = -2 \log \int L(h)\pi(h|v^2, h_{-1}, h_0)dh + 2 \times 3 + (-2) \log (\text{likelihood of } \hat{q}(x)) + 2 \times (\text{the number of free parameters of } \hat{q}(x)) . \tag{2.14}$$

Note that the third and fourth terms on the right-hand side in (2.14) correspond to the AIC for a reference distribution $\hat{q}(x)$ (see the Appendix for the reasoning which leads to this expression of ABIC). If the value of this ABIC is larger than that of AIC for the reference distribution, that is, the first and second terms of the right-hand side in (2.14) take a nonnegative value, then we adopt the estimate given by the model with the minimum AIC as our final estimate.

2.5 *Behavior of the estimation procedure*

To demonstrate the behavior of the present procedure, we shall consider the following experiment. Two probability distributions are defined as

$$f_1^*(x) = \frac{1}{2}\varphi(x-5) + \frac{1}{2}\varphi(x-9) , \tag{2.15}$$

$$f_2^*(x) = \frac{1}{2}\varphi(x-5) + \frac{1}{2}\varphi(x-8) , \tag{2.16}$$

where $\varphi(x-m)$ denotes the density function of the normal distribution with mean m and variance 1. From each distribution, random samples of size 200 were generated and the Kullback-Leibler information quantity (K-L information quantity) between the true density and the final estimate obtained by the present procedure was calculated. From these operations, repeated 100 times for each case, we obtained 0.0283 and 0.0266 as the respective averages of the K-L information quantities. On the other hand, GALTHY gave 0.0342 and 0.0306 for the same data sets. Although these values indicate the comparative superiority of our approach over that of GALTHY, the difference of the K-L information quantities between the two procedures is not so large when we take into consideration the sample size of 200. For example, the difference between 200×0.0283 and 200×0.0342 equals -1.18 which is not so large as to prove a significant difference in the goodness of fit to the data (see, Sakamoto *et al.* (1986), pp. 84–85). However, there is a remarkable gap in the stability of estimates between these two procedures. The solid line in Fig. 1 shows the distribution

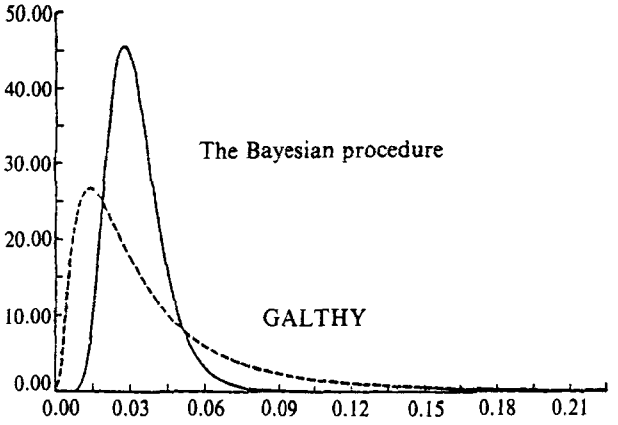

Fig. 1. Density functions estimated from the K-L information quantities.

of 100 estimates of the K-L information quantities, each of which was calculated from the final estimates obtained by the present procedure for the first distribution in this experiment. The corresponding density function for GALTHY is shown with a dotted line in the same figure. From this figure, we can see that the density function for GALTHY has a longer tail than that for our procedure. This illustrates that the estimates by GALTHY have a larger variance than the present procedure. These observations imply that our procedure based on a Bayesian model produces moderate estimates compared with GALTHY, which is based on an ordinary parametric model.

For more numerical examples of the present procedure, see Ishiguro and Sakamoto (1984).

3. A Bayesian approach to statistical test problems

3.1 *Goodness of fit test problem*

The chi-square goodness of fit test has been widely used to evaluate the fitting of theoretical distributions to observations. The test statistic is based on the squares of differences between observed and hypothetical frequencies falling into properly fixed classes. This implies that this test procedure is interpreted as the comparison between constrained and unconstrained multinomial models (Sakamoto *et al.* (1986)) and that it involves some serious problems. For example, the decision depends significantly on the method of initial classification. The procedure proposed in the preceding section is not seriously affected by this problem and provides an effective approach to the goodness of fit test problem.

Suppose that we wish to test the null hypothesis that a set of data was drawn from a distribution having a density function $f_0(x)$. If we assign $f_0(x)$ for $q(x)$ in the Step 1 in Subsection 2.4, this test can be viewed as the test

of the hypothesis $r(y) = 1$ $(0 \leq y \leq 1)$ in the expression

$$f(x) = r(y)q(x) , \tag{3.1}$$

against the alternative $r(y) \neq 1$. To deal with this problem, we assume the following models.

$$\text{MODEL(0)}: \quad \pi(h) = \delta(h - (1,\ldots,1)) , \tag{3.2}$$

and

$$\text{MODEL(1)}: \quad \pi(h) = \pi(h|v^2, h_{-1}, h_0) , \tag{3.3}$$

where $\pi(h)$ denotes the prior distribution of h, $\delta(h - (1,\ldots,1))$ is the delta function with mass concentrated at $h = (1,\ldots,1)$, and $\pi(h|v^2, h_{-1}, h_0)$ is defined by (2.7). From the relation

$$\int \left\{ \prod_{i=1}^{n} r(y_i|h) \right\} \pi(h)dh = \int \left\{ \prod_{i=1}^{n} r(y_i|h) \right\} \delta(h - (1,\ldots,1))dh$$

$$= \prod_{i=1}^{n} r(y_i|h = (1,\ldots,1)) = \prod_{j=1}^{C} \left\{ \frac{C \exp(1/C)}{\sum_{k=1}^{C} \exp(1/C)} \right\}^{n_j} = 1 ,$$

we can see that under MODEL(0), the ABIC for $f(x)$ degenerates to the AIC for $q(x)$. Thus, if $f_0(x)$ is, for example, a normal distribution, the statistic ABIC(0) to evaluate the goodness of fit of the MODEL(0) is given by

$$\text{ABIC(0)} = \text{AIC(0)} = n \log 2\pi + n \log \hat{\sigma}^2 + n + 4 , \tag{3.4}$$

where $\hat{\sigma}^2$ denotes the maximum likelihood estimate of variance. It is clear that the ABIC for MODEL(1) is given by

$$\begin{aligned} \text{ABIC(1)} = (-2) \log g(y|\hat{v}^2, \hat{h}_{-1}, \hat{h}_0) + 2 \times 3 \\ + \{n \log 2\pi + n \log \hat{\sigma}^2 + n + 4\} , \end{aligned} \tag{3.5}$$

where

$$g(y|v^2, h_{-1}, h_0) = \int L(h)\pi(h|v^2, h_{-1}, h_0)dh . \tag{3.6}$$

Here $L(h)$ is defined by (2.6) and $g(y|\hat{v}^2, \hat{h}_{-1}, \hat{h}_0)$ denotes the model which attains the minimum ABIC. Note that the AIC for the reference distribution is commonly included in these two ABIC's.

The performance of this procedure is illustrated by the following experiment. When the true distribution $f^*(x)$ is $0.5\varphi(x - m_0) + 0.5\varphi(x - (m_0 - 4))$, the frequencies at which ABIC(1) takes a smaller value than AIC(0) are 172 out of 500 times for $n = 50$, 427 out of 500 times for $n = 100$, and 100 out of 100 times for $n = 200$. Although the performance of the procedure depends on the sample size and the complexity in shape of the true distribution, our procedure seems to serve for most practical uses as the sample size increases. Note that the present procedure frees the data analyst from arbitrariness in the choice of an initial classification.

Of course, GALTHY has shown another approach to the goodness of fit test problem. However, GALTHY does not provide the estimate of measure for the goodness of fit of its final estimate, or of the final model itself, and the goodness of the final estimate cannot be compared with that of estimates obtained by any other models. This problem can be solved by the consistent use of the statistic ABIC, or AIC.

3.2 *Two-sample "nonparametric" test problem*

3.2.1 *Model and procedure*

For simplicity, two sets of samples $y_1 = (y_1^{(1)}, \ldots, y_{n_1}^{(1)})$ and $y_2 = (y_1^{(2)}, \ldots, y_{n_2}^{(2)})$ are temporarily assumed to be included in the interval $[0, 1]$. We assume that those data sets are drawn from two populations having unknown probability density functions $r(y|h^{(1)})$ and $r(y|h^{(2)})$, respectively. Suppose that we wish to test the hypothesis

$$H_0: \quad h^{(1)} = h^{(2)}, \tag{3.7}$$

against

$$H_1: \quad h^{(1)} \neq h^{(2)}. \tag{3.8}$$

To deal with this problem we assume the following models:

$$\text{MODEL(0)}: \quad \pi(h^{(1)}, h^{(2)}) = \pi(h^{(1)})\delta(h^{(1)} - h^{(2)}), \tag{3.9}$$

and

$$\text{MODEL(1)}: \quad \pi(h^{(1)}, h^{(2)}) = \pi(h^{(1)})\pi(h^{(2)}), \tag{3.10}$$

where the prior distribution is defined by

$$\begin{aligned} \pi(h^{(l)}) &\equiv \pi(h^{(l)}|v_{(l)}^2, h_{-1}^{(l)}, h_0^{(l)}) \\ &= \prod_{j=1}^{c} \frac{1}{\sqrt{2\pi}\, v_{(l)}} \exp\left\{ -\frac{1}{2v_{(l)}^2} (h_j^{(l)} - 2h_{j-1}^{(l)} + h_{j-2}^{(l)})^2 \right\}, \end{aligned} \tag{3.11}$$

$$l = 1, 2 .$$

MODEL(0) means that its parameter space degenerates to the sub-space $\{(h^{(1)}, h^{(2)}) | h^{(1)} = h^{(2)} = h^{(0)}\}$. Under MODEL(0) the likelihood of the Bayesian model is given by

$$\begin{aligned}(3.12)\quad &\int\int \left\{ \prod_{i=1}^{n_1} r(y_i^{(1)}|h^{(1)}) \right\} \left\{ \prod_{k=1}^{n_2} r(y_k^{(2)}|h^{(2)}) \right\} \pi(h^{(1)}, h^{(2)}) dh^{(1)} dh^{(2)} \\ &= \int\int \left\{ \prod_{i=1}^{n_1} r(y_i^{(1)}|h^{(1)}) \right\} \left\{ \prod_{k=1}^{n_2} r(y_k^{(2)}|h^{(2)}) \right\} \pi(h^{(1)}) \delta(h^{(1)} - h^{(2)}) dh^{(1)} dh^{(2)} \\ &= \int \left\{ \prod_{i=1}^{n_1} r(y_i^{(1)}|h^{(0)}) \right\} \left\{ \prod_{k=1}^{n_2} r(y_k^{(2)}|h^{(0)}) \right\} \pi(h^{(0)}) dh^{(0)} .\end{aligned}$$

Thus, the ABIC(0) for the MODEL(0) is given by

$$\begin{aligned}(3.13)\quad &\text{ABIC(0)} \\ &= (-2) \log \int \left\{ \prod_{i=1}^{n_1} r(y_i^{(1)}|h^{(0)}) \right\} \left\{ \prod_{k=1}^{n_2} r(y_k^{(2)}|h^{(0)}) \right\} \pi(h^{(0)}) dh^{(0)} + 2 \times 3 \\ &\equiv (-2) \log g^{(0)}(y|\hat{v}_{(0)}^2, \hat{h}_{-1}^{(0)}, \hat{h}_0^{(0)}) + 2 \times 3 .\end{aligned}$$

Of course, $\pi(h^{(0)})$ in these expressions is defined by putting $l = 0$ in (3.11).

On the other hand, ABIC(1) for the MODEL(1) is given by

$$\begin{aligned}(3.14)\quad \text{ABIC(1)} &= (-2) \log g^{(1)}(y|\hat{v}_{(1)}^2, \hat{h}_{-1}^{(1)}, \hat{h}_0^{(1)}) + 2 \times 3 \\ &\quad + (-2) \log g^{(2)}(y|\hat{v}_{(2)}^2, \hat{h}_{-1}^{(2)}, \hat{h}_0^{(2)}) + 2 \times 3 ,\end{aligned}$$

where

$$\begin{aligned}(3.15)\quad &g^{(l)}(y|v_{(l)}^2, h_{-1}^{(l)}, h_0^{(l)}) \\ &= \int \left\{ \prod_{i=1}^{n_l} r(y_i^{(l)}|h^{(l)}) \right\} \pi(h^{(l)}|v_{(l)}^2, h_{-1}^{(l)}, h_0^{(l)}) dh^{(l)} , \qquad l = 1, 2 .\end{aligned}$$

Here the likelihood for each data set is defined by

$$(3.16)\quad \prod_{i=1}^{n_l} r(y_i^{(l)}|h^{(l)}) = \prod_{j=1}^{C} \left\{ \frac{C \exp(h_j^{(l)}/C)}{\sum_{k=1}^{C} \exp(h_k^{(l)}/C)} \right\}^{n_j^{(l)}} , \qquad l = 1, 2 .$$

As seen from the discussions in the preceding sections, this procedure can be easily extended to the general case where each observation does not

necessarily belong to the closed interval $[0, 1]$. Suppose that we have two sets of data $x_i^{(1)}$, $i = 1,\ldots,n_1$ and $x_k^{(2)}$, $k = 1,\ldots,n_2$ and we denote the reference distributions chosen by the minimum AIC procedure for the first, second and pooled data sets by $\hat{q}_{(1)}(x)$, $\hat{q}_{(2)}(x)$ and $\hat{q}_{(0)}$, respectively. By a reasoning similar to that in (2.14) or (3.5), we define the ABIC's for this case as follows:

$$\begin{aligned}(3.17)\qquad \mathrm{ABIC}(0) &= (-2)\log g^{(0)}(y|\hat{v}_{(0)}^2, \hat{h}_{-1}^{(0)}, \hat{h}_0^{(0)}) + 2\times 3 \\ &\quad + (-2)\log(\text{likelihood of } \hat{q}_{(0)}(x)) \\ &\quad + 2\times(\text{the number of parameters of } \hat{q}_{(0)}),\end{aligned}$$

$$\begin{aligned}(3.18)\qquad \mathrm{ABIC}(1) &= (-2)\log g^{(1)}(y|\hat{v}_{(1)}^2, \hat{h}_{-1}^{(1)}, \hat{h}_0^{(1)}) + 2\times 3 \\ &\quad + (-2)\log(\text{likelihood of } \hat{q}_{(1)}(x)) \\ &\quad + 2\times(\text{the number of parameters of } \hat{q}_{(1)}) \\ &\quad + (-2)\log g^{(2)}(y|\hat{v}_{(2)}^2, \hat{h}_{-1}^{(2)}, \hat{h}_0^{(2)}) + 2\times 3 \\ &\quad + (-2)\log(\text{likelihood of } \hat{q}_{(2)}(x)) \\ &\quad + 2\times(\text{the number of parameters of } \hat{q}_{(2)}).\end{aligned}$$

As seen from (3.12), MODEL(0) shows the fitting of a particular model to the pooled data. On the other hand, MODEL(1) shows the fitting of different models to every data set; its ABIC is defined as the sum of the ABIC's for each data set. This procedure can be easily extended to the case of the k-sample "nonparametric" test problems. If we have, for example, data drawn from three populations, we have only to assume five possible models and choose the model with the minimum ABIC value from among them. Also, in this case, the necessary ABIC is defined as the sum of the ABIC's, if not all the population distributions are assumed to be homogeneous.

3.2.2 *Numerical examples*

The first example is two sets of data concerning the glucose concentration measured by the reference and test methods, each of which consists of 46 measurements (Leurgans (1980), pp. 216–217). For this data set, our procedure has given ABIC(0) = 1026.71 and ABIC(1) = 1044.81. Clearly, ABIC(0) is smaller than ABIC(1). This result indicates that the two underlying distributions are not distinguishable. Each data set follows an identical distribution which is shown in Fig. 2.

We shall next turn our attention to measurements of the sepal length of the flowers of fifty plants from each of the three species, Iris setosa (denoted by G1), Iris versicolor (G2) and Iris virginica (G3), which are given in Table 1 in Fisher (1936). We tried to check the homogeneity of the population distributions of these three species. For this data set, we

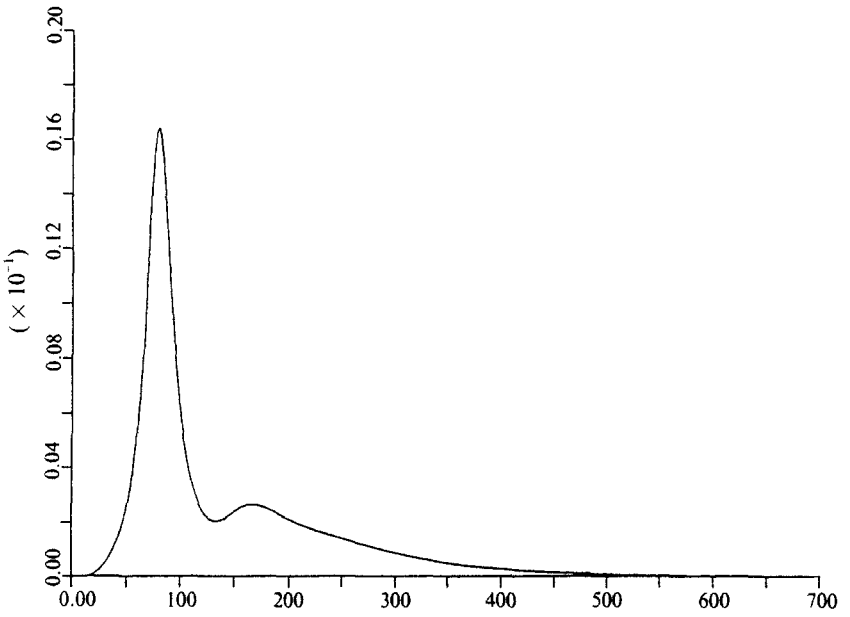

Fig. 2. Density function estimated from glucose data.

obtained the results shown in Table 1. This table shows five possible models rearranged in increasing order of the value of ABIC. In this table, for example, the sign (G1, G2 + G3) in the second row represents the model which assumes that G2 and G3 have an identical population distribution, but that G1 has another distribution. Since a model with a smaller ABIC value is expected to be a better model, this result suggests that the best model is (G1, G2, G3), which assumes that the respective distributions of the three species are mutually different.

These examples may not persuade the reader of the practical utility of our procedure since their true density functions are unknown. We shall clarify the behavior of our procedure by the following experiment.

Suppose that the first and second density functions $f^{(1)}$ and $f^{(2)}$ are identical and are defined by

Table 1. Analysis of Fisher's Iris data.

Rank	MODEL	ABIC
1	(G1, G2, G3)	218.77
2	(G1, G2 + G3)	243.26
3	(G1 + G2, G3)	294.04
4	(G1 + G3, G2)	348.89
5	(G1 + G2 + G3)	368.00

$$f^{(1)}(x) = f^{(2)}(x) = \frac{1}{2}\varphi(x-8) + \frac{1}{2}\varphi(x-12), \tag{3.19}$$

and that the third one is defined by

$$f^{(3)}(x) = \frac{1}{\sigma}\varphi((x-10)/\sigma), \tag{3.20}$$

where $\sigma^2 = 5$ and $f^{(3)}(x)$ means the density function of the normal density distribution with mean 10 and variance σ^2. The dotted line in Fig. 3 shows the density function for the former, and the solid line, that of the latter. Clearly, the density functions (3.19) and (3.20) have equal means and variances but different shapes. We generated random samples of size 100 from each distribution and checked whether our procedure detected the true structure, i.e., that the density functions $f^{(1)}$ and $f^{(2)}$ are identical, but that $f^{(3)}$ is different from $f^{(1)}$ or $f^{(2)}$. The tests were repeated 100 times and the results shown in Table 2 were obtained. The abbreviation for the models in this table is the same as in Table 1. In this table, an empty cell represents the zero frequency. This table shows that our procedure detected the true structure in 98 out of 99 cases in which the K-L information quantity between an estimated and the true density functions took the minimum value. This means that the decision by our procedure for this

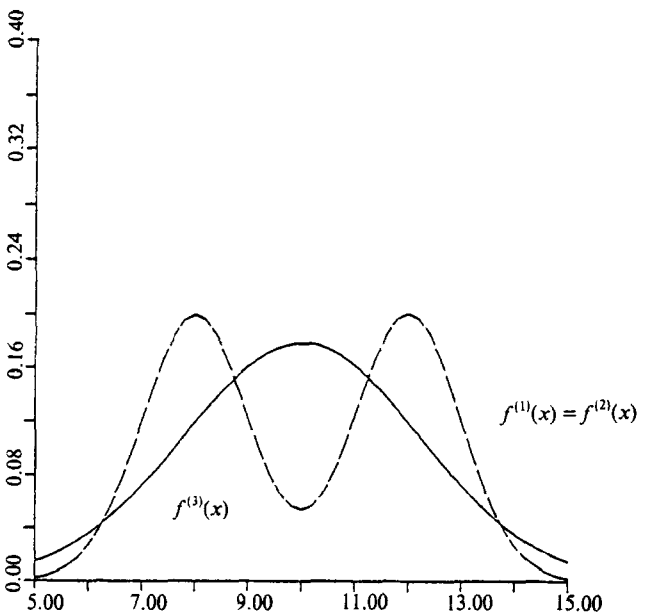

Fig. 3. $f^{(1)}(x) = f^{(2)}(x) = \frac{1}{2}\varphi(x-8) + \frac{1}{2}\varphi(x-12)$, $f^{(3)}(x) = \frac{1}{\sigma}\varphi((x-10)/\sigma)$, $\sigma^2 = 5$.

Table 2. Frequencies of the minimum of the K-L information and ABIC value.

Frequency of minimum ABIC	Frequency of minimum K-L information				
	(G1, G2, G3)	(G1, G2 + G3)	(G1 + G3, G2)	(G1 + G2, G3)	(G1 + G2 + G3)
(G1, G2, G3)					
(G1, G2 + G3)					
(G1 + G3, G2)					
(G1 + G2, G3)	1			98	
(G1 + G2 + G3)				1	

experiment has coincided very often with the indication of the K-L information. In addition, the average of these 100 ABIC values has taken the minimum value at the true structure. It is clear that the performance of the procedure depends on the structure and the sample size. However, this results is a typical example which illustrates the behavior of the present procedure and seems to imply that our procedure is practical if the sample size is sufficiently large.

4. Conclusion

The advantage of the classical test procedures is that they are very simple to handle if we ignore the problems that the conventional test procedures do not have grounds for a reasonable determination of a level of significance and that it is practically impossible to deal with multiple hypothesis situations. On the other hand, the advantage of our procedure is that it can evaluate the goodness of fit of each model over all the assumed models and that it can provide the estimate of the underlying population distribution, if necessary.

Recently, we have reviewed standard statistical procedures from a consistent viewpoint, namely, the construction of a parametric model and its evaluation by the Akaike information criterion, AIC (Sakamoto *et al.* (1986)). This viewpoint frees our method from individual theories of sampling distributions and various statistical tables. However, alternative approaches to the "nonparametric" test problems have not been included in the book since such problems can rarely be solved by ordinary parametric models. The procedure proposed in the present paper shows the feasibility of a consistent parametric approach to problems which have conventionally been dealt with by the "nonparametric" test procedures.

Acknowledgements

The authors are grateful to Professor H. Akaike and Professor G. Kitagawa for their helpful comments and to Mr. K. Katsura of the

Institute of Statistical Mathematics for his computing assistance.

Appendix

Suppose that a true density function $f^*(x)$ of original data x and its estimate $\hat{f}(x)$ are expressed as

$$f^*(x) = r^*(\hat{Q}(x))\hat{q}(x) ,$$

and

$$\hat{f}(x) = \int \hat{r}(\hat{Q}(x)|h)\pi(h)\hat{q}(x)dh ,$$

respectively. We then evaluate the goodness of fit of $\hat{f}(x)$ to $f^*(x)$ by

$$\begin{aligned}
&\int f^*(x) \log \hat{f}(x)dx \\
&\quad = \int f^*(x) \log \int \hat{r}(\hat{Q}(x)|h)\pi(h)dh\hat{q}(x)dx \\
&\quad = \int f^*(x) \log \int \hat{r}(\hat{Q}(x)|h)\pi(h)dhdx + \int f^*(x) \log \hat{q}(x)dx \\
&\quad = \int r^*(\hat{Q}(x))\hat{q}(x) \log \int \hat{r}(\hat{Q}(x)|h)\pi(h)dhdx + \int f^*(x) \log \hat{q}(x)dx .
\end{aligned}$$

Using $y = \hat{Q}(x)$ and $dy = \hat{q}(x)dx$, we have that

$$\int r^*(\hat{Q}(x))\hat{q}(x) \log \int \hat{r}(\hat{Q}(x)|h)\pi(h)dhdx = \int r^*(y) \log \int \hat{r}(y|h)\pi(h)dhdy .$$

Hence, we finally obtain

$$\int f^*(x) \log \hat{f}(x)dx = \int r^*(y) \log \int \hat{r}(y|h)\pi(h)dhdy + \int f^*(x) \log \hat{q}(x)dx .$$

which shows that the goodness of fit of $\hat{f}(x)$ can be evaluated by the sum of the two terms measuring the goodness of fit of $\hat{r}(y)$ and that of $\hat{q}(x)$.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd Internat. Symp. Inform. Theory*, (eds. B. N. Petrov and F. Csaki), 267–806, Akademiai Kiado, Budapest.

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

Akaike, H. (1980). Likelihood and Bayes procedure, *Bayesian Statistics*, (eds. J. M. Bernardo, M. H. DeGroot, D. U. Lindley and A. F. M. Smith), University Press, Valencia, Spain.

Akaike, H. and Arahata, E. (1978). GALTHY, a probability density estimation, *Comput. Sci. Monographs*, No. 9, The Institute of Statistical Mathematics, Tokyo.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics*, **7**, 179–188.

Ishiguro, M. and Sakamoto, Y. (1984). A Bayesian approach to the probability density estimation, *Ann. Inst. Statist. Math.*, **36**, 523–538.

Leurgans, S. (1980). Evaluating laboratory measurement techniques, *Biostatistics Casebook*, (eds. R. G. Miller Jr., B. Efron, B. W. M. Brown Jr. and L. E. Moses), 190–219, Wiley, New York.

Neyman, J. (1937). 'Smooth' test for goodness of fit, *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*, Reidel, Dordrecht, Holland.