Biometrika (1996), **83**, 4, *pp.* 875–890
Printed in Great Britain

Generalised information criteria in model selection

BY SADANORI KONISHI

Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812, Japan

AND GENSHIRO KITAGAWA

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

SUMMARY

The problem of evaluating the goodness of statistical models is investigated from an information-theoretic point of view. Information criteria are proposed for evaluating models constructed by various estimation procedures when the specified family of probability distributions does not contain the distribution generating the data. The proposed criteria are applied to the evaluation of models estimated by maximum likelihood, robust, penalised likelihood, Bayes procedures, etc. We also discuss the use of the bootstrap in model evaluation problems and present a variance reduction technique in the bootstrap simulation.

Some key words: AIC; Bayes approach; Efficient bootstrap simulation; Information criterion; *M*-estimators; Penalised likelihood; Predictive distribution; Statistical functional.

1. INTRODUCTION

Akaike's information criterion AIC (Akaike, 1973, 1974) is a model evaluation and selection tool which can be thought of in terms of estimating the Kullback & Leibler (1951) information of the true model with respect to the fitted model. It can be derived under the assumptions that

(i) estimation is by maximum likelihood, and

(ii) this is carried out in a parametric family of distributions including the true model.

We introduce criteria which estimate the same Kullback–Leibler information as does the AIC, relaxing both the assumptions (i) and (ii). The essential point, in our work and Akaike's, is to approximate the bias in a very simple but clearly deficient estimate of the Kullback–Leibler information in question. In § 2 we develop the general theory, using functional-type estimators, and show the simplifications resulting from assumption (i) alone, and then from both (i) and (ii). In § 3 the information criteria proposed are applied to the evaluation of the various types of models based on robust, penalised likelihood and Bayes estimation procedures. In § 4 the variance reduction technique in the bootstrap simulation is investigated, and it is shown that the variance of the bootstrap estimates caused by simulation can be reduced without any analytical derivations. The derivation of the proposed criteria is given in the Appendix.

2. GENERALISED INFORMATION CRITERIA

Suppose that X_n is a random sample of size n from an unknown distribution $G(x)$ having probability density function $g(x)$. Estimation is done within a parametric family of distributions with densities $\{f(x\,|\,\theta);\ \theta\in\Theta\}$, which may or may not contain g. Here θ is a p-dimensional vector of unknown parameters. The predictive density function $f(z\,|\,\hat{\theta})$ for a future observation z can be constructed by replacing the unknown parameter vector θ by some efficient estimate $\hat{\theta}$, for which maximum likelihood, robust or penalised likelihood procedures are used for parameter estimation. The development in terms of this simple predictive density will also apply, with some modification, to the Bayesian one.

Model selection can be approached by trying to make small the Kullback–Leibler information (Kullback & Leibler, 1951)

$$I\{g(z);\ f(z\,|\,\hat{\theta})\}:=\int g(z)\log g(z)\,dz-\int g(z)\log f(z\,|\,\hat{\theta})\,dz, \tag{1}$$

which is a measure of the divergence of $g(z)$ relative to $f(z\,|\,\hat{\theta})$. The first term on the right-hand side of (1) does not depend on the model and only the second term, the expected log likelihood,

$$\eta(X_n;\ G):=\int g(z)\log f(z\,|\,\hat{\theta})\,dz=\int \log f(z\,|\,\hat{\theta})\,dG(z) \tag{2}$$

is relevant.

A simple estimator of $\eta(X_n;\ G)$ is given by substituting for G the empirical distribution $\hat{G}$, yielding the log likelihood

$$\eta(X_n;\ \hat{G})=\frac{1}{n}\sum_{\alpha=1}^{n}\log f(X_\alpha\,|\,\hat{\theta}). \tag{3}$$

Ordinarily this provides an overestimate of the expected log likelihood, since $\hat{G}$ corresponds more closely to $\hat{\theta}$ than does the true G. Writing the bias as

$$b(G):=E_G\{\eta(X_n;\ \hat{G})-\eta(X_n;\ G)\},$$

we define an information criterion based on the bias-corrected log likelihood as

$$\frac{1}{n}\sum_{\alpha=1}^{n}\log f(X_\alpha\,|\,\hat{\theta})-b(\hat{G}).$$

Akaike (1973) showed that $b(G)$ is asymptotically p/n under the assumptions that estimation is by maximum likelihood and the specified parametric model contains the true distribution g. Hence the AIC can be considered as a criterion obtained by correcting the asymptotic bias of the log likelihood of the estimated model.

In order to remove both of the assumptions imposed on AIC, we first consider bias correction of the log likelihood when $\hat{\theta}$ is a functional estimator; namely when $\hat{\theta}$ is defined by $\hat{\theta}=T(\hat{G})$ with T being a suitably defined p-dimensional regular functional. It is assumed that the functional under consideration is Fisher consistent, that is $T(F_\theta)=\theta$ for all $\theta\in\Theta$, where F_θ is the distribution function of the specified probability density $f(x\,|\,\theta)$. Then the bias of the log likelihood is given in the following theorem.

THEOREM 2·1. *Let $\eta(X_n;\ G)$ and $\eta(X_n;\ \hat{G})$ be the expected log likelihood in* (2) *and the log likelihood in* (3), *respectively. Suppose that the parametric family of distributions does not necessarily contain the true distribution and that the statistical functional $T(.)$ is second-order compact differentiable at G. Then the asymptotic bias of the log likelihood in the* *estimation of the expected log likelihood is given by*

$$E_G\{\eta(X_n;\,\hat{G})-\eta(X_n;\,G)\}=\frac{1}{n}b_1(G)+o\left(\frac{1}{n}\right), \tag{4}$$

where

$$b_1(G)=\operatorname{tr}\left\{\int T^{(1)}(z;\,G)\left.\frac{\partial\log f(z\,|\,\theta)}{\partial\theta'}\right|_{T(G)}dG(z)\right\}, \tag{5}$$

and $T^{(1)}(z;\,G)=(T_1^{(1)}(z;\,G),\ldots,T_p^{(1)}(z;\,G))'$ *is the influence function of a p-dimensional functional* $T(G)$ *at the distribution* G.

The derivation is given in the Appendix.

Let $b_1(\hat{G})$ be a bias estimate obtained by replacing the unknown distribution G by the empirical distribution $\hat{G}$. Then we have an information criterion based on the bias corrected log likelihood as follows:

$$\begin{aligned}\text{GIC}\,(X_n;\,\hat{G})&=-2n\left\{\frac{1}{n}\sum_{\alpha=1}^{n}\log f(X_\alpha\,|\,\hat{\theta})-\frac{1}{n}b_1(\hat{G})\right\}\\&=-2\sum_{\alpha=1}^{n}\log f(X_\alpha\,|\,\hat{\theta})+\frac{2}{n}\sum_{\alpha=1}^{n}\operatorname{tr}\left\{T^{(1)}(X_\alpha;\,\hat{G})\left.\frac{\partial\log f(X_\alpha\,|\,\theta)}{\partial\theta'}\right|_{\hat{\theta}}\right\},\end{aligned} \tag{6}$$

where $T^{(1)}(X_\alpha;\,\hat{G})=(T_1^{(1)}(X_\alpha;\,\hat{G}),\ldots,T_p^{(1)}(X_\alpha;\,\hat{G}))'$ is the p-dimensional empirical influence function defined by

$$T_i^{(1)}(X_\alpha;\,\hat{G})=\lim_{\varepsilon\to 0}\,[T_i\{(1-\varepsilon)\hat{G}+\varepsilon\delta(X_\alpha)\}-T_i(\hat{G})]/\varepsilon,$$

with $\delta(X_\alpha)$ being the point mass at X_α. The influence function $T^{(1)}(z;\,G)$ and its estimate $T^{(1)}(X_\alpha;\,\hat{G})$ are respectively the derivatives of $T(G)$ and $T(\hat{G})$ with respect to the probability measures $\delta(z)$ and $\delta(X_\alpha)$. For the problem of choosing among different models, we select the model for which the value of the information criterion GIC $(X_n;\,\hat{G})$ is smallest.

Theorem 2·1 produces in a simple way criteria for evaluating the models estimated by various types of procedures. We now show how the results simplify for the case of maximum likelihood estimation.

The maximum likelihood estimator, $\hat{\theta}_{\text{ML}}$, can be expressed as $\hat{\theta}_{\text{ML}}=T_{\text{ML}}(\hat{G})$, where T_{ML} is the p-dimensional functional given as a solution of the implicit equation

$$\int\left.\frac{\partial}{\partial\theta}\log f(z\,|\,\theta)\right|_{T_{\text{ML}}(G)}dG(z)=0.$$

Replacing G by $(1-\varepsilon)G+\varepsilon\delta(z)$ in the above equation and differentiating with respect to ε, we have the p-dimensional influence function of $\hat{\theta}_{\text{ML}}=T_{\text{ML}}(\hat{G})$ in the form

$$T_{\text{ML}}^{(1)}(z;\,G)=J(G)^{-1}\left.\frac{\partial\log f(z\,|\,\theta)}{\partial\theta}\right|_{T_{\text{ML}}(G)}, \tag{7}$$

where

$$J(G)=-\int\left.\frac{\partial^2\log f(z\,|\,\theta)}{\partial\theta\,\partial\theta'}\right|_{T_{\text{ML}}(G)}dG(z). \tag{8}$$

Substituting the influence function $T^{(1)}_{\mathrm{ML}}$ given by (7) in the result (5) yields the asymptotic bias $\operatorname{tr}\{J(G)^{-1}I(G)\}$, where

$$I(G)=\int \frac{\partial \log f(z\,|\,\theta)}{\partial\theta}\frac{\partial \log f(z\,|\,\theta)}{\partial\theta'}\bigg|_{T_{\mathrm{ML}}(G)} dG(z). \qquad (9)$$

Thus for the case of maximum likelihood estimation the information criterion GIC defined by (6) reduces to

$$\mathrm{TIC}\,(X_n;\,\hat{G})\text{:}= -2\sum_{\alpha=1}^{n}\log f(X_\alpha\,|\,\hat{\theta}_{\mathrm{ML}})+2\operatorname{tr}\{J(\hat{G})^{-1}I(\hat{G})\}, \qquad (10)$$

where $J(G)$ and $I(G)$ are respectively given by (8) and (9). This criterion was originally introduced by Takeuchi (1976) and discussed by Stone (1977) and Shibata (1989).

If the specified parametric family of densities contains the true distribution, that is $g(z)=f(z\,|\,\theta_0)$, or $G=F_{\theta_0}$, for some θ_0 in Θ, then we see that $\operatorname{tr}\{J(F_{\theta_0})^{-1}I(F_{\theta_0})\}=p$, noting the well-known identity $I(F_{\theta_0})=J(F_{\theta_0})$. Hence TIC in (10) is further reduced to

$$\mathrm{AIC}= -2\sum_{\alpha=1}^{n}\log f(X_\alpha\,|\,\hat{\theta}_{\mathrm{ML}})+2p, \qquad (11)$$

the criterion proposed by Akaike (1973, 1974). Stone (1977) showed that AIC is asymptotically equivalent to the cross-validation which is a naive method for estimating prediction error.

We have so far concentrated on deriving information criteria for the fitted densities in the context of functional statistics. The generalised information criterion given by (6) may not directly be applied to the evaluation of the models with the Bayes and nonfunctional estimators. However, if the Bayesian predictive density, $h(z\,|\,X_n)$, and the model $f(z\,|\,\hat{\theta}_U)$ with nonfunctional estimator $\hat{\theta}_U$ may be approximated by $f(z\,|\,\hat{\theta})+O_p(n^{-1})$ with functional statistic $\hat{\theta}=T(\hat{G})$, then an information criterion similar to GIC in (6) can be obtained as follows.

THEOREM 2·2. *Let $h(z\,|\,X_n)$ be a model fitted to the data X_n drawn from the distribution $G(x)$. Assume that the fitted model can be expressed as $h(z\,|\,X_n)=f(z\,|\,\hat{\theta})+O_p(n^{-1})$, where $\hat{\theta}=T(\hat{G})$ is a functional statistic for functional $T(G)=(T_1(G),\ldots,T_p(G))'$. Then an information criterion for the evaluation of the model $h(z\,|\,X_n)$ is given by*

$$-2\sum_{\alpha=1}^{n}\log h(X_\alpha\,|\,X_n)+\frac{2}{n}\sum_{\alpha=1}^{n}\operatorname{tr}\left\{T^{(1)}(X_\alpha;\,\hat{G})\frac{\partial \log f(X_\alpha\,|\,\theta)}{\partial\theta'}\bigg|_{\hat{\theta}}\right\},$$

where $T^{(1)}(X_\alpha;\,\hat{G})=(T^{(1)}_1(X_\alpha;\,\hat{G}),\ldots,T^{(1)}_p(X_\alpha;\,\hat{G}))'$ is the empirical influence function of $\hat{\theta}=T(\hat{G})$.

If the estimator $\hat{\theta}_U$ differs from a functional statistic $\hat{\theta}$ by an error of order $O_p(n^{-1})$, then $f(z\,|\,\hat{\theta}_U)=f(z\,|\,\hat{\theta})+O_p(n^{-1})$, so that Theorem 2·2 can be directly applied to the evaluation of $f(z\,|\,\hat{\theta}_U)$. For example, the usual unbiased variance estimator may be expressed as

$$T(\hat{G})\{1+n^{-1}+O_p(n^{-2})\},$$

where $T(G)=2^{-1}\int\int(x-y)^2\,dG(x)\,dG(y)$. We note that GIC defined by (6) is still valid with the log likelihood of a fitted model $h(z\,|\,X_n)$ if this predictive distribution can be approximated as $f(z\,|\,\hat{\theta})+O_p(n^{-1})$ with functional statistic $\hat{\theta}=T(\hat{G})$.

The criterion AIC given by(11) has some attractive properties in practical applications;

the bias correction term does not require any analytical derivation, and it can be applied in an automatic way in various situations (Kitagawa, 19**8**7). A number of successful applications of AIC in statistical data analysis have been reported; see, e.g., Bozdogan (1994), Sakamoto, Ishiguro & Kitagawa (1986). The problem is whether the bias of the log likelihood of a fitted model is well approximated by the number of parameters when a specified model deviates from the true distribution.

Figure 1 plots the true bias $nb(G)$ and the estimated bias $\mathrm{tr}\{J(\hat{G})^{-1}I(\hat{G})\}$, with standard errors, in addition to the asymptotic bias $\mathrm{tr}\{J(G)^{-1}I(G)\}$, for sample sizes $n=25$, 100, 400 and 1600. The true density function $g(x)$ and the model $f(z\,|\,\theta)$ are respectively assumed to be

$$g(x)=(1-\varepsilon)\phi(x\,|\,0,1)+\varepsilon\phi(x\,|\,0,d^2) \quad (d\gg 1), \quad f(x\,|\,\theta)=\phi(x\,|\,\mu,\sigma^2), \tag{12}$$

where $\theta=(\mu,\sigma^2)$, and $\phi(x\,|\,\mu,\sigma^2)$ denotes the density function of the normal distribution with mean μ and variance σ^2. The variance d^2 is arbitrarily set to 10. For this situation the values of biases are plotted for mixing proportion ε ranging from 0 to 0·2, in which the

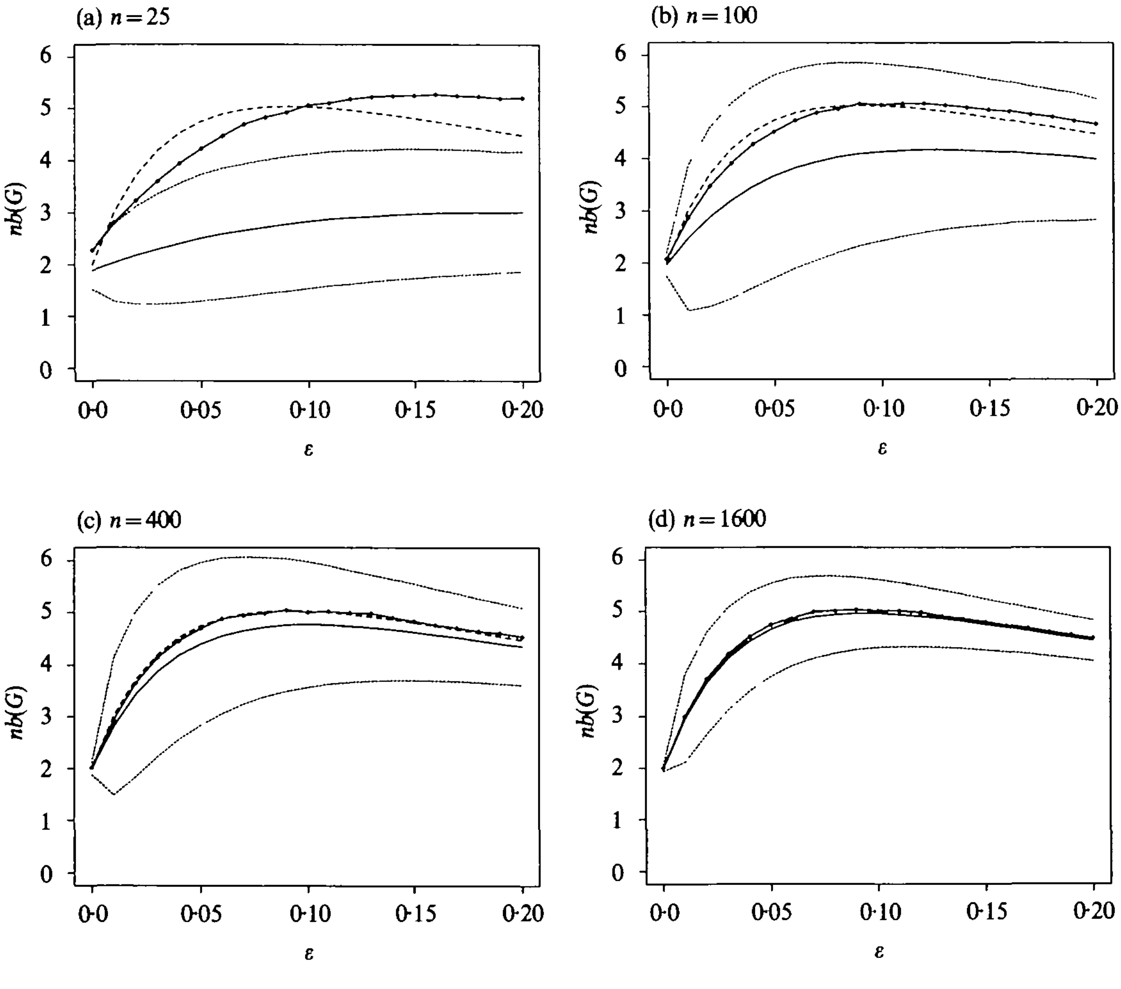

Fig. 1. Comparison of the true bias $n \times b(G)$ (•–•), the asymptotic bias $\mathrm{tr}\{J(G)^{-1}I(G)\}$(- - - -) and the estimated asymptotic bias $\mathrm{tr}\{J(\hat{G})^{-1}I(\hat{G})\}$ (——) with standard errors (·····) for the sample sizes $n=25$, 100, 400 and 1600.

quantities other than the asymptotic bias are evaluated by a Monte Carlo simulation with 100 000 repetitions.

It may be seen from Fig. 1 that the log likelihood of a fitted model has a significant bias as an estimate of the expected log likelihood and that the bias is significantly larger than 2, the approximation by AIC, if the mixing proportion ε is around 0·07. It should be noticed here that the amount of contamination in the observed data is often between 5% and 7% and this is the region where the bias is large. We observe that for large sample sizes the true bias, the asymptotic bias and the estimated asymptotic bias coincide quite well. On the other hand, for smaller sample sizes such as $n=25$, the estimated asymptotic bias underestimates the true asymptotic bias and the true bias.

3. CRITERIA FOR EVALUATING VARIOUS STATISTICAL MODELS

3·1. *General*

This section gives criteria to evaluate the goodness of the models estimated by robust and penalised maximum likelihood procedures. We also consider an information criterion for the predictive densities of Bayesian methods.

3·2. *Robust estimation*

Let $f(x\,|\,\hat{\theta}_M)$ be the fitted model to the data X_n from the true distribution $G(x)$, where $\hat{\theta}_M$ is an M-estimator defined as the solution of the implicit equations

$$\sum_{\alpha=1}^{n} \psi_i(X_\alpha, \hat{\theta}_M) = 0 \quad (i=1,\ldots,p), \tag{13}$$

with ψ_i a function on $\mathscr{X} \times \Theta$ $(\Theta \subset R^p)$ (Huber, 1981). The solution $\hat{\theta}_M$ can be written as $T_M(\hat{G})$ for the functional given by

$$\int \psi_i\{z, T_M(G)\}\, dG(z) = 0 \quad (i=1,\ldots,p). \tag{14}$$

It is known, see e.g. Hampel et al. (1986, p. 230), that the influence function of the M-estimator is

$$T_M^{(1)}(z;\, G) = M(\psi, G)^{-1}\psi\{z, T_M(G)\},$$

where $\psi=(\psi_1,\ldots,\psi_p)'$ and the $p\times p$ matrix $M(\psi, G)$ is given by

$$M(\psi, G)' = -\int \left.\frac{\partial\psi(z,\theta)'}{\partial\theta}\right|_{T_M(G)} dG(z), \tag{15}$$

under the assumption that the matrix M is nonsingular. Then it follows from Theorem 2·1 that the asymptotic bias of the log likelihood $n^{-1}\sum \log f(X_\alpha\,|\,\hat{\theta}_M)$ in the estimation of the expected log likelihood (2) with $\hat{\theta}_M$ is

$$b_M^{(1)}(G) = \operatorname{tr}\left[M(\psi, G)^{-1} \int \psi\{z, T(G)\} \left.\frac{\partial \log f(z\,|\,\theta)}{\partial\theta'}\right|_{T_M(G)} dG(z)\right], \tag{16}$$

which is estimated by $b_M^{(1)}(\hat{G})$.

Let us now consider the situation that the parametric family of densities $\{f(x\,|\,\theta);\ \theta\in\Theta\subset R^p\}$ contains the true density, that is $g(x)=f(x\,|\,\theta_0)$ for some θ_0 in Θ.

Assume that the functional T_M is Fisher consistent. Then the matrix M in (15) can be expressed as

$$M(\psi, F_\theta) = \int \psi(z, \theta) \frac{\partial \log f(z \mid \theta)}{\partial \theta'} dF_\theta(z),$$

where $F_\theta(x)$ is the distribution function of $f(x \mid \theta)$ (Hampel et al., 1986, p. 231). By replacing $M(\psi, G)$ in (16) by $M(\psi, F_\theta)$, we have $b_M^{(1)}(F_\theta) = p$, which does not depend on unknown parameters. The results on information criteria for M-estimator are summarised in the following theorem.

THEOREM 3·1. *Let $f(x \mid \hat{\theta}_M)$ be the estimated model based on the M-estimator $\hat{\theta}_M$ defined by* (13). *Then an information criterion for $f(x \mid \hat{\theta}_M)$ is*

$$\text{IC}_M(X_n; \hat{G}) := -2 \sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\theta}_M) + 2b_M^{(1)}(\hat{G}), \tag{17}$$

where $b_M^{(1)}(G)$ is given by (16). *If the specified model contains the true distribution and the functional T_M defined by* (14) *is Fisher consistent, then the information criterion* (17) *can be reduced to*

$$-2 \sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\theta}_M) + 2p,$$

where p is the number of parameters within the model $f(x \mid \theta)$.

It might be noted that the maximum likelihood estimator is an M-estimator, corresponding to

$$\psi\{x, T_M(G)\} = \partial \log f(x \mid \theta)/\partial\theta|_{T_M(G)}.$$

Hence TIC given by (10) can be also obtained by taking this $\psi\{x, T_M(G)\}$ in Theorem 3·1. The use of Theorem 3·1 is illustrated through the following examples.

Example 3·1. Let $\{(y_i, x_i);\ i = 1, \ldots, n\}$ $(y_i \in R, x_i \in R^p)$ be a sample of independent, identically distributed random variables with common distribution $G(y, x)$ having density $g(y, x)$. Consider the linear model

$$y_i = x_i'\theta + \varepsilon_i \quad (i = 1, \ldots, n), \tag{18}$$

where θ is a p-dimensional parameter vector. Let $F(y, x \mid \theta)$ be a model distribution with density $f(y, x \mid \theta) = f_1(y - x'\theta)f_2(x)$, in which the error ε_i is assumed to be independent of x_i and a scale parameter is ignored. In the linear model (18), the M-estimator is given as the solution of the system of equations

$$\sum_{i=1}^{n} \psi(y_i - x_i'\hat{\theta}_R)x_i = 0. \tag{19}$$

The influence function of the M-estimator defined by (19) at the distribution G is

$$T_R^{(1)}(G) = \left[\int \psi'\{y - x'T_R(G)\}xx'\, dG\right]^{-1} \psi\{y - x'T_R(G)\}x,$$

where $\psi'(z) = \partial\psi(z)/\partial z$ and $T_R(G)$ is the functional given by

$$\int \psi\{y - x'T_R(G)\}x\, dG = 0.$$

Then it follows from Theorem 3·1 that the asymptotic bias of the log likelihood of $f(y, x \mid \hat{\theta}_R)$ is

$$b_R^{(1)}(G) = \mathrm{tr}\left(\left[\int \psi'\{y - x'T_R(G)\}xx'\,dG\right]^{-1} \int \psi\{y - x'T_R(G)\}x \left.\frac{\partial \log f(y, x \mid \theta)}{\partial \theta'}\right|_{T_R(G)} dG\right). \tag{20}$$

Suppose that the true density g can be written in the form $g(y, x) = g_1(y - x'\theta)g_2(x)$, and that the M-estimator defined by (19) is the maximum likelihood estimator for the model $f(y, x \mid \theta)$, that is $\partial \log f(y, x \mid \theta)/\partial\theta = \psi(y - x'\theta)x$. Then the asymptotic bias $b_R^{(1)}(G)$ in (20) can be reduced to $E_{g_1}(\psi')^{-1}E_{g_1}(\psi^2)p$, which agrees with the result given by Ronchetti (1985, p. 23).

For an ordinary Gaussian linear regression model, Sugiura (1978) obtained the bias of a log likelihood in closed form when the specified model includes the true one, and Hurvich & Tsai (1991) considered this problem under the assumption that the candidate models do not include the true model. In regression Mallow's (1973) C_p is a classical model selection procedure and equivalent to AIC. A robust version of Mallow's C_p was presented by Ronchetti & Staudte (1994).

Example 3·2: *Numerical result.* Consider the parametric model $F_\theta(x) = \Phi\{(x - \mu)/\sigma\}$, where Φ is the standard normal distribution function. It is assumed that the parametric family of distributions $\{F_\theta(x);\ \theta \in \Theta \subset R^2\}$ contains the true distribution generating the data X_n. The location and scale parameters are respectively estimated by the median, $\hat{\mu}_m = \mathrm{med}_i\{X_i\}$, and the median absolute deviation, $\hat{\sigma}_m = (1/c)\,\mathrm{med}_i\{|X_i - \mathrm{med}_j(X_j)|\}$, where $c = \Phi^{-1}(0{\cdot}75)$ is chosen to make $\hat{\sigma}_m$ Fisher consistent for Φ.

Table 1. *Estimates of the biases for the M-estimator and the maximum likelihood estimator obtained by a Monte Carlo simulation*

	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1600$
Robust	3·839	2·569	2·250	2·125	2·056	2·029	2·013
MLE	2·229	2·079	2·047	2·032	2·014	2·002	2·003

Table 1 compares the biases ($\times n$) of the log likelihoods for the M-estimator $(\hat{\mu}_m, \hat{\sigma}_m)$ and the maximum likelihood estimator $(\hat{\mu}, \hat{\sigma}^2)$ obtained by averaging over 100 000 repeated Monte Carlo trials. Note that the bias for the maximum likelihood estimator is analytically given by $nb(G) = 2n/(n - 3)$. From the table it may be observed that both of the biases actually converge to the asymptotic bias, 2, as the sample size n gets larger and that the convergence of the bias for the robust estimator is slower than that for the maximum likelihood estimator.

3·3. *Penalised likelihood estimation*

We consider a criterion for models estimated by procedures based on penalised likelihoods (Good & Gaskins, 1971) in the context of M-estimators. Let T_λ be the p-dimensional statistical functional determined by the system of equations

$$\int \left.\frac{\partial}{\partial\theta}\{\log f(z \mid \theta) - \lambda k(\theta)\}\right|_{T_\lambda(G)} dG(z) = 0,$$

where the penalty function $k(\theta)$ is assumed to be twice continuously differentiable. Then the estimate $\hat{\theta}_\lambda = T_\lambda(\hat{G})$ is a maximum penalised likelihood estimate defined by the functional T_λ.

Taking

$$\psi\{z, T_\lambda(G)\} = \partial\,\{\log f(z\,|\,\theta) - \lambda k(\theta)\}/\partial\theta$$

at $\theta = T_\lambda(G)$ in the formula (16) and replacing G by the empirical distribution function $\hat{G}$ give the information criterion for the model $f(x\,|\,\hat{\theta}_\lambda)$ as

$$\mathrm{IC}_P(X_n;\, \hat{G}) = -2 \sum_{\alpha=1}^{n} \log f(X_\alpha\,|\,\hat{\theta}_\lambda) + 2\,\mathrm{tr}\,\{J_\lambda(\hat{G})^{-1} I_\lambda(\hat{G})\}, \tag{21}$$

where

$$I_\lambda(\hat{G}) = \frac{1}{n}\sum_{\alpha=1}^{n} \psi(X_\alpha, \hat{\theta}_\lambda)\, \frac{\partial \log f(X_\alpha\,|\,\theta)}{\partial\theta'}\bigg|_{\hat{\theta}_\lambda}, \quad J_\lambda(\hat{G}) = -\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial\psi(X_\alpha, \theta)'}{\partial\theta}\bigg|_{\hat{\theta}_\lambda}.$$

An optimal value of λ is chosen such that the information criterion IC_P in (21) is minimal. If the penalty function $k(\theta)$ does not depend on the distribution G, then $\int \psi\{z, T_\lambda(G)\}\,\partial k(\theta)/\partial\theta'\, dG(z) = 0$, so that the information criterion (21) agrees with the result obtained by Shibata (1989). Green (1987) examined penalised likelihood estimation in the context of general regression models and introduced various types of roughness penalties.

3·4. *Predictive distributions*

The predictive density function in the Bayesian framework is defined by

$$h(z\,|\,X_n) = \int f(z\,|\,\theta)\pi(\theta\,|\,X_n)\, d\theta,$$

where $\pi(\theta\,|\,X_n)$ is the posterior density function for θ based on a prior $\pi(\theta)$ and the data X_n. It is required to approximate the predictive density with an error of order at least $O_p(n^{-1})$, for which we use Laplace's methods for integrals in the Bayesian framework developed by Tierney & Kadane (1986) and Davison (1986). The predictive density may be expressed as

$$\begin{aligned} h(z\,|\,X_n) &= \int f(z\,|\,\theta) \prod_{\alpha=1}^{n} f(X_\alpha\,|\,\theta)\pi(\theta)\, d\theta \bigg/ \int \prod_{\alpha=1}^{n} f(X_\alpha\,|\,\theta)\pi(\theta)\, d\theta \\ &= \int \exp\{q(\theta\,|\,X_n) + \log f(z\,|\,\theta)\}\, d\theta \bigg/ \int \exp\{q(\theta\,|\,X_n)\}\, d\theta, \end{aligned}$$

where $q(\theta\,|\,X_n) = \sum \log f(X_\alpha\,|\,\theta) + \log \pi(\theta)$.

Let $\hat{\theta}_\pi$ and $\hat{\theta}_\pi(z)$ be the solutions of the equations

$$\partial\{q(\theta\,|\,X_n)\}/\partial\theta = 0, \quad \partial\{q(\theta\,|\,X_n) + \log f(z\,|\,\theta)\}/\partial\theta = 0,$$

respectively. Then it follows from the expression given by Davison (1986, p. 325) that the Laplace approximation to the predictive density is

$$h(z\,|\,X_n) = \{|J_q|/|J_q(z)|\}^{\frac{1}{2}} \exp[q\{\hat{\theta}_\pi(z)\,|\,X_n\} - q(\hat{\theta}_\pi\,|\,X_n) + \log f\{z\,|\,\hat{\theta}_\pi(z)\}]\,\{1 + O_p(n^{-2})\},$$

where

$$J_q = -\partial^2\{q(\theta\,|\,X_n)\}/\partial\theta\,\partial\theta'|_{\hat{\theta}_\pi}, \quad J_q(z) = -\partial^2\{q(\theta\,|\,X_n) + \log f(z\,|\,\theta)\}/\partial\theta\,\partial\theta'|_{\hat{\theta}_\pi(z)}.$$

Noting that $\hat{\theta}_\pi - \hat{\theta}_{\mathrm{ML}} = O_p(n^{-1})$ and $\hat{\theta}_\pi(z) - \hat{\theta}_{\mathrm{ML}} = O_p(n^{-1})$ for the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$, and using the result $|J_q|/|J_q(z)| = 1 + O_p(n^{-1})$ produce the approximation $h(z\,|\,X_n) = f(z\,|\,\hat{\theta}_{\mathrm{ML}}) + O_p(n^{-1})$.

Hence it follows from Theorem 2·2 that a criterion for the evaluation of the predictive density $h(z\,|\,X_n)$ is given by

$$-2\sum_{\alpha=1}^{n} \log h(X_\alpha\,|\,X_n) + 2\,\mathrm{tr}\{J(\hat{G})^{-1} I(\hat{G})\},$$

where $J(G)$ and $I(G)$ are defined by (8) and (9), respectively.

4. NUMERICAL APPROACH

4·1. *General objectives*

It may be seen that the bias correction for the log likelihood of a predictive density in the estimation of the expected log likelihood is essential to construct an information criterion. The bootstrap methods (Efron, 1979) offer an alternative approach to estimate the bias of the log likelihood numerically. The use of the bootstrap in a simple situation was introduced by Efron (1986, p. 467). Ishiguro, Sakamoto & Kitagawa (1996) examined through numerical examples the performance of the bootstrap bias corrected information criterion called EIC. This section discusses the use of the bootstrap in statistical model evaluation problems and introduces the variance reduction procedure in bootstrap simulation.

4·2. *Bootstrap bias estimates*

Let $h(z\,|\,X_n)$ be a model fitted to the data X_n drawn from the true distribution $G(x)$. The model $h(z\,|\,X_n)$ may be constructed in various ways based on the specified parametric family of densities $\{f(x\,|\,\theta);\ \theta \in \Theta\}$. Conditional on the observed data X_n, let X_n^* be the bootstrap sample generated according to the empirical distribution $\hat{G}$ of X_n. The bootstrap analogue of the expected log likelihood of $h(z\,|\,X_n)$ defined by (2) is

$$\eta(X_n^*;\ \hat{G}) := \int \log h(z\,|\,X_n^*)\, d\hat{G}(z) = \frac{1}{n}\sum_{\alpha=1}^{n} \log h(X_\alpha\,|\,X_n^*).$$

By noting that the same sample X_n^* is used to estimate $\eta(X_n^*;\ \hat{G})$, the log likelihood of $h(z\,|\,X_n^*)$ is constructed as

$$\eta(X_n^*;\ \hat{G}^*) := \int \log h(z\,|\,X_n^*)\, d\hat{G}^*(z) = \frac{1}{n}\sum_{\alpha=1}^{n} \log h(X_\alpha^*\,|\,X_n^*),$$

where $\hat{G}^*$ is the empirical distribution function of the bootstrap sample X_n^*. Then the bootstrap bias of the log likelihood is given by $\hat{b}_B(\hat{G}) := E_{\hat{G}}\{\eta(X_n^*;\ \hat{G}^*) - \eta(X_n^*;\ \hat{G})\}$, the bootstrap estimate of $b(G) := E_G\{\eta(X_n;\ \hat{G}) - \eta(X_n;\ G)\}$.

The bias $\hat{b}_B(\hat{G})$ is approximated by averaging $\eta(X_n^*;\ \hat{G}^*) - \eta(X_n^*;\ \hat{G})$ over a large number of repeated bootstrap samples. By subtracting the bootstrap bias estimate from the log likelihood, we have

$$\mathrm{EIC}\,(X_n;\ \hat{G}) := -2\sum_{\alpha=1}^{n} \log h(X_\alpha\,|\,X_n) + 2n\hat{b}_B(\hat{G}).$$

Suppose the fitted model may be expressed as $h(z\,|\,X_n) = f(z\,|\,\hat{\theta}) + O_p(n^{-1})$, where $\hat{\theta}$ is a

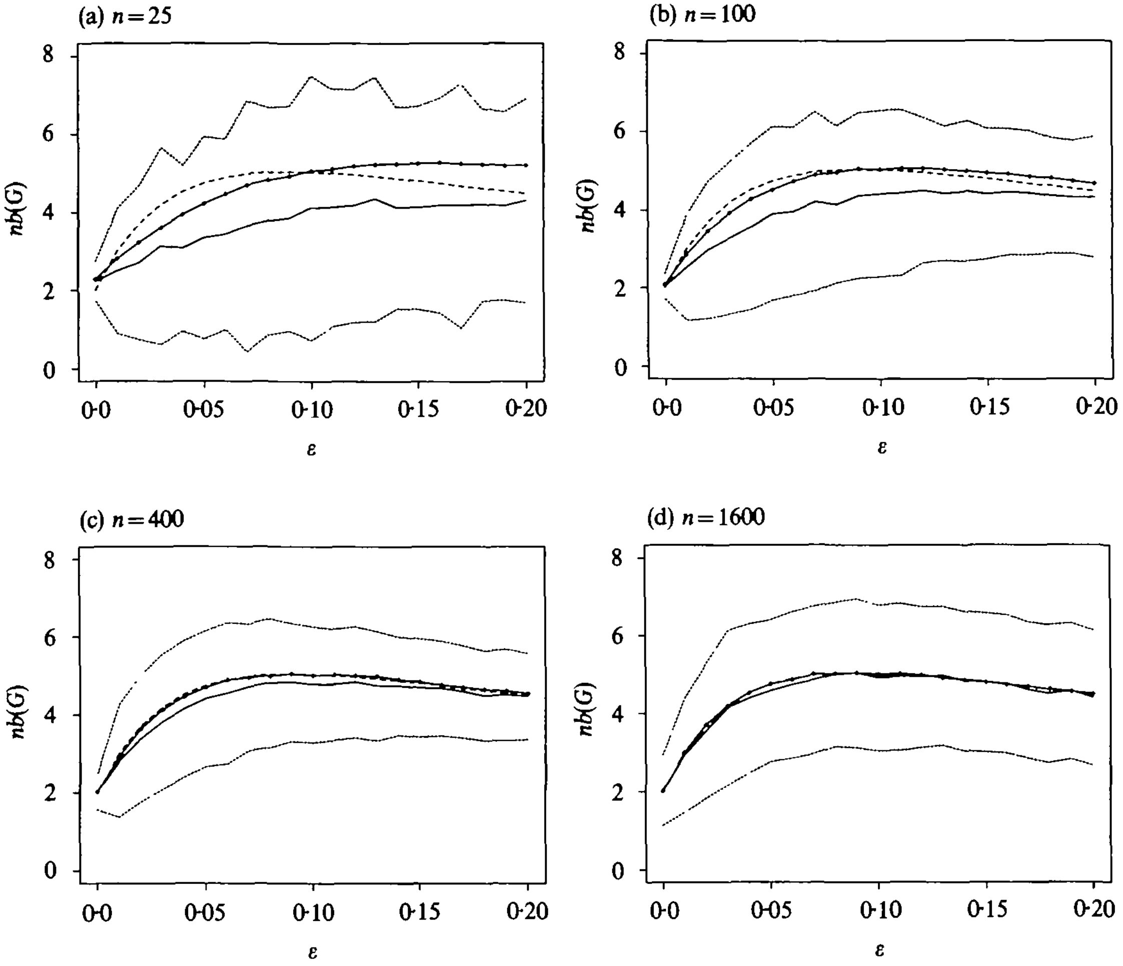

Fig. 2. Comparison of the true bias $n \times b(G)$ (•-•), the asymptotic bias $\mathrm{tr}\{J(G)^{-1}I(G)\}$(- - -) and the bootstrap estimate of the bias $n \times b_B(\hat{G})$ (——) with standard errors (·····) for the sample sizes $n = 25$, 100, 400 and 1600.

functional statistic. Then it follows from the expressions (4) and (5) in Theorem 2·1 that the bootstrap bias estimate $n\hat{b}_B(\hat{G})$ converges in probability to $b_1(G)$ in (4) as n tends to infinity.

Figure 2 shows the bootstrap estimates ($\times\, n$) of the biases of the log likelihood for the cases given in Fig. 1. In Monte Carlo experiments, 1000 repeated random samples were generated from a mixture of two normal populations in (12) for the sample sizes $n = 25$, 100, 400 and 1600, and 1000 bootstrap replications were taken for each trial. From Fig. 2 we see that the means of the bootstrap estimates are close to the true biases $nb(G)$, for large sample sizes. It may be seen from Figs 1 and 2 that the standard deviations of the bootstrap estimates are comparable to those of $\mathrm{tr}\{J(\hat{G})^{-1}I(\hat{G})\}$ for $n = 100$ and 400, but are larger than those for $n = 25$ and 1600. The reason for the increase of variance in large sample case and the reduction of variance in the bootstrap simulation are discussed in § 4·3.

4·3. *Variance reduction in bootstrap simulation*

The bootstrap bias estimates are usually approximated by a Monte Carlo simulation. The bias estimates obtained numerically include both the randomness of the observed data and simulation error which decreases as the number of bootstrap replication increases. It will be shown that the variance associated with a Monte Carlo simulation can be reduced automatically without any analytical argument.

Let $T(X_n; G) = \eta(X_n; \hat{G}) - \eta(X_n; G)$ for the log likelihood of $f(z \mid \hat{\theta})$ in (3) and the expected log likelihood in (2). Then the bias of the log likelihood $\eta(X_n; \hat{G})$ or equivalently the expectation of $T(X_n; G)$ may be decomposed into three parts

$$E_G\{T(X_n; G)\} = E_G\{T_1(X_n; G) + T_2(X_n; G) + T_3(X_n; G)\}, \tag{22}$$

where

$$\begin{aligned} T_1(X_n; G) &= \frac{1}{n} \sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\theta}) - \frac{1}{n} \sum_{\alpha=1}^{n} \log f\{X_\alpha \mid T(G)\}, \\ T_2(X_n; G) &= \frac{1}{n} \sum_{\alpha=1}^{n} \log f\{X_\alpha \mid T(G)\} - \int g(z) \log f\{z \mid T(G)\}\, dz, \\ T_3(X_n; G) &= \int g(z) \log f\{z \mid T(G)\}\, dz - \int g(z) \log f(z \mid \hat{\theta})\, dz. \end{aligned} \tag{23}$$

We examined the bootstrap estimates of

$$nE_G\{T_1(X_n; G)\}, \quad nE_G\{T_2(X_n; G)\}, \quad nE_G\{T_3(X_n; G)\}$$

and their variances in the bootstrap simulation through Monte Carlo experiments. It is assumed that the true distribution and the specified model are both normal, and the model is estimated by the maximum likelihood methods. The simulation results were obtained by averaging over 1000 repeated Monte Carlo trials. For each Monte Carlo trial $B = 100$ bootstrap samples are taken and variances are calculated based on 100 bootstrap bias estimates. Table 2 shows the bootstrap estimates and variances for sample sizes $n = 25$, 100, 400 and 1600, in which the notation BT$_1$, BT$_2$, BT$_3$, BT and BT$_{13}$ refers to the simulation results for $nE_G\{T_1(X_n; G)\}$, $nE_G\{T_2(X_n; G)\}$, $nE_G\{T_3(X_n; G)\}$, $nE_G\{T(X_n; G)\}$ and $nE_G\{T_1(X_n; G) + T_3(X_n; G)\}$, respectively.

We observe from the experiment that the bootstrap estimate of the second term BT$_2$ is almost zero and does not affect the estimate of the total expected value in (22) for each case, but its variance is extremely large, especially for $n = 400$ or $n = 1600$. In contrast the bootstrap estimates BT$_1$ and BT$_3$ have much smaller variances. This clearly shows the reason for the increase of variances of the bootstrap estimates in large sample cases.

The simulation results suggest eliminating the second term $T_2(X_n; G)$ and applying the

Table 2. *Simulation results: The bootstrap estimates,* BSE, *and variances,* Var, *for* BT$_1 = nE_G\{T_1(X_n; G)\}$, BT$_2 = nE_G\{T_2(X_n; G)\}$, BT$_3 = nE_G\{T_3(X_n; G)\}$, BT $= nE_G\{T(X_n; G)\}$ *and* BT$_{13} = nE_G\{T_1(X_n; G) + T_3(X_n; G)\}$

	$n = 25$		$n = 100$		$n = 400$		$n = 1600$	
	BSE	Var	BSE	Var	BSE	Var	BSE	Var
BT$_1$	1·230	0·036	1·041	0·013	1·009	0·011	1·003	0·010
BT$_2$	−0·001	0·110	0·002	0·480	0·011	1·962	0·008	7·904
BT$_3$	1·002	0·012	0·995	0·010	0·999	0·010	1·000	0·010
BT	2·232	0·242	2·038	0·564	2·019	2·034	2·010	7·974
BT$_{13}$	2·232	0·086	2·036	0·047	2·008	0·041	2·003	0·040

bootstrap to approximate the bias $E_G\{T(X_n; G)\}$ in (22). Then the bootstrap estimate of $E_G\{T_1(X_n; G)+T_3(X_n; G)\}$ is the same as that of $E_G\{T(X_n; G)\}$, but the variance is much smaller. This implies that, if the bootstrap technique is applied to the modified formula, the number of bootstrap replications may be greatly reduced.

The theoretical justification for use of the simple variance reduction technique mentioned above is as follows. By an argument similar to that in Appendix, $T(X_n; G)$ in (22) can be expanded as

$$T(X_n; G)=\frac{1}{n}\sum_{\alpha=1}^{n} \text{IF}\,(X_\alpha; G)+O_p\left(\frac{1}{n}\right),$$

where

$$\text{IF}\,(X_\alpha; G):=\log f\{X_\alpha \mid T(G)\} - \int g(z)\log f\{z \mid T(G)\}\, dz,$$

the influence function of $T(X_n; G)$. Noting that $n^{-1}\sum \text{IF}\,(X_\alpha; G)=T_2(X_n; G)$ in (23) and $E_G\{\text{IF}\,(X_\alpha; G)\}=0$, the bootstrap estimate of the bias (22) can be rewritten as

$$\begin{aligned} E_{\hat{G}}\{T(X_n^*; \hat{G})\} &= E_{\hat{G}}\left\{T(X_n^*; \hat{G})-\frac{1}{n}\sum_{\alpha=1}^{n} \text{IF}\,(X_\alpha^*; \hat{G})\right\} \\ &= E_{\hat{G}}\left\{\frac{1}{n}\sum_{\alpha=1}^{n} \log f(X_\alpha^* \mid \hat{\theta}^*)-\frac{1}{n}\sum_{\alpha=1}^{n} \log f(X_\alpha^* \mid \hat{\theta})\right. \\ &\qquad \left. +\frac{1}{n}\sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\theta})-\frac{1}{n}\sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\theta}^*)\right\}, \end{aligned} \qquad (24)$$

where $\hat{\theta}^*=T(\hat{G}^*)$ for the empirical distribution function $\hat{G}^*$ of X_n^*.

It can be easily shown that the asymptotic conditional variances of

$$B^{-1}\sum T(X_n^*; \hat{G}), \quad B^{-1}\sum\{T(X_n^*; \hat{G})-n^{-1}\sum \text{IF}\,(X_\alpha^*; \hat{G})\}$$

are of order $B^{-1}n^{-1}$ and $B^{-1}n^{-2}$, respectively, where B is the number of bootstrap replications. This implies that the use of the decomposition in (24) is superior to the ordinary bootstrap bias estimate in the sense that the simulation error is of smaller order.

The method discussed above is essentially the same as the linear approximation method investigated by Davison, Hinkley & Schechtman (1986). The advantage in the proposed variance reduction method is that it can be applied automatically in a wide variety of situations without any knowledge of influence functions for estimators, whereas the linear approximation method requires formulae for linear terms.

5. Discussion

The main aim of this paper was to provide criteria which enable us to evaluate various types of statistical models. We derived the criteria from the information-theoretic point of view in the context of functional statistics without the assumption that the specified family of probability distributions contains the true structure generating the data. An advantage of the criteria proposed is that they are not restricted to maximum likelihood estimation but may be applied to the evaluation of models constructed by a variety of procedures as discussed in § 3. It might be noted that the existing model evaluation criteria such as AIC, TIC in (10) and Schwarz's (1978) Bayes information criterion cover only models estimated by the maximum likelihood methods.

The criteria obtained in this paper were constructed by correcting the bias of a log likelihood in the estimation of the expected log likelihood of a fitted density. Ideally we would like to obtain a criterion, IC $(X_n; \hat{G})$, which satisfies the condition that

$$E_G\{\text{IC}\,(X_n; \hat{G}) - \eta(X_n; G)\} = O(n^{-j})$$

for larger values of j, where $\eta(X_n; G)$ is the expected log likelihood given by (2). The bias-corrected version of the log likelihood is second-order correct or accurate for $\eta(X_n; G)$ in the sense that the remainder is of order $O(n^{-2})$ in the above condition. We obtained a criterion which achieves third-order accuracy in the context of functional statistics. The result is however very lengthy and seems to be of no practical use. By bootstrapping the bias-corrected log likelihood, we may obtain a criterion with third-order accuracy. The large variances of bootstrap estimates should however be examined carefully. Further work remains to be done towards constructing more refined results, especially for small sample sizes.

Acknowledgement

The authors would like to thank the Editor, Professor Dawid, the Associate Editor and a referee for their helpful comments and suggestions which improved the quality of this paper considerably.

Appendix

Derivation of Theorem 2·1

We recall that an estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$ is a functional, for which there exists a p-dimensional functional T such that $\hat{\theta} = T(\hat{G}) = (T_1(\hat{G}), \ldots, T_p(\hat{G}))'$. Suppose that T_i is second-order compact differentiable at G for $i = 1, \ldots, p$. Then the functional Taylor series expansion for $\hat{\theta}_i = T_i(\hat{G})$ is, up to order n^{-1},

$$\hat{\theta}_i = T_i(G) + \frac{1}{n}\sum_{\alpha=1}^{n} T_i^{(1)}(X_\alpha; G) + \frac{1}{n^2}\frac{1}{2}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} T_i^{(2)}(X_\alpha, X_\beta; G) + o_p(n^{-1}), \tag{A1}$$

where $T_i^{(\alpha)}$ are defined as symmetric functions such that, for an arbitrary distribution H,

$$\frac{d^\alpha}{d\varepsilon^\alpha} T_i\{(1-\varepsilon)G + \varepsilon H\} = \int \ldots \int T_i^{(\alpha)}(x_1, \ldots, x_\alpha; G) \prod_{j=1}^{\alpha} d\{H(x_j) - G(x_j)\}$$

at $\varepsilon = 0$ and

$$\int T_i^{(\alpha)}(x_1, \ldots, x_\alpha; G)\, dG(x_j) = 0 \quad (1 \leqslant j \leqslant \alpha).$$

Then it may be seen that, as $n \to +\infty$, $n^{\frac{1}{2}}\{\hat{\theta} - T(G)\}$ is asymptotically normally distributed with mean vector 0 and covariance matrix $\Sigma = (\sigma_{ij})$, where

$$\sigma_{ij} = \int T_i^{(1)}(z; G) T_j^{(1)}(z; G)\, dG(z).$$

Also it can be readily shown that the asymptotic bias of $\hat{\theta}$ for $T(G)$ is $E_G\{\hat{\theta} - T(G)\} = b/n + o(n^{-1})$, where $b = (b_1, \ldots, b_p)'$ with

$$b_i = \frac{1}{2}\int T_i^{(2)}(z, z; G)\, dG(z).$$

Expanding $\log f(z|\hat{\theta})$ in a Taylor series around $\hat{\theta}=T(G)$ and substituting (A1) in the resulting expansion, we have stochastic expansions for the expected log likelihood $\eta(X_n, G)=\int g(z)\log f(z|\hat{\theta})\,dz$ and the log likelihood $\eta(X_n;\hat{G})=(1/n)\sum\log f(X_\alpha|\hat{\theta})$ as follows:

$$\begin{aligned}\eta(X_n; G) = {}& \int g(z)\log f\{z|T(G)\}\,dz + \frac{1}{n}\sum_{i=1}^{p}\sum_{\alpha=1}^{n} T_i^{(1)}(X_\alpha; G)\int g(z)\left.\frac{\partial \log f(z|\theta)}{\partial\theta_i}\right|_{T(G)} dz \\ & + \frac{1}{2n^2}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\left\{\sum_{i=1}^{p} T_i^{(2)}(X_\alpha, X_\beta; G)\int g(z)\left.\frac{\partial \log f(z|\theta)}{\partial\theta_i}\right|_{T(G)} dz\right. \\ & \left. + \sum_{i=1}^{p}\sum_{j=1}^{p} T_i^{(1)}(X_\alpha; G)T_j^{(1)}(X_\beta; G)\int g(z)\left.\frac{\partial^2 \log f(z|\theta)}{\partial\theta_i\,\partial\theta_j}\right|_{T(G)} dz\right\} + o_p\left(\frac{1}{n}\right), \\ \eta(X_n; \hat{G}) = {}& \frac{1}{n}\sum_{\alpha=1}^{n}\log f\{X_\alpha|T(G)\} + \frac{1}{n^2}\sum_{i=1}^{p}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} T_i^{(1)}(X_\alpha; G)\left.\frac{\partial \log f(X_\beta|\theta)}{\partial\theta_i}\right|_{T(G)} \\ & + \frac{1}{2n^3}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\sum_{\gamma=1}^{n}\left\{\sum_{i=1}^{p} T_i^{(2)}(X_\alpha, X_\beta; G)\left.\frac{\partial \log f(X_\gamma|\theta)}{\partial\theta_i}\right|_{T(G)}\right. \\ & \left. + \sum_{i=1}^{p}\sum_{j=1}^{p} T_i^{(1)}(X_\alpha; G)T_j^{(1)}(X_\beta; G)\left.\frac{\partial^2 \log f(X_\gamma|\theta)}{\partial\theta_i\,\partial\theta_j}\right|_{T(G)}\right\} + o_p\left(\frac{1}{n}\right).\end{aligned}$$

Taking expectations term by term yields

$$E_G\{\eta(X_n; G)\} = \int g(z)\log f\{z|T(G)\}\,dz + \frac{1}{n}\left[b'\kappa - \frac{1}{2}\operatorname{tr}\{\Sigma(G)J(G)\}\right] + o\left(\frac{1}{n}\right), \tag{A2}$$

$$\begin{aligned}E_G\{\eta(X_n; \hat{G})\} = {}& \int g(z)\log f\{z|T(G)\}\,dz + \frac{1}{n}\left[\sum_{i=1}^{p}\int T_i^{(1)}(z; G)\left.\frac{\partial \log f(z|\theta)}{\partial\theta_i}\right|_{T(G)} dG(z)\right. \\ & \left. + b'\kappa - \frac{1}{2}\operatorname{tr}\{\Sigma(G)J(G)\}\right] + o\left(\frac{1}{n}\right),\end{aligned} \tag{A3}$$

where κ and $J(G)$ are given by

$$\kappa = \int\left.\frac{\partial \log f(z|\theta)}{\partial\theta}\right|_{T(G)} dG(z), \quad J(G) = -\int\left.\frac{\partial^2 \log f(z|\theta)}{\partial\theta\,\partial\theta'}\right|_{T(G)} dG(z).$$

This shows immediately that the asymptotic bias of the log likelihood is given by the result (5) in Theorem 2·1. By an argument similar to that discussed above, we have the result given in Theorem 2·2.

We have derived the terms of order $1/n^2$ in (A2) and (A3) under the fourth-order compact differentiability condition on T. The results are, however, very lengthy and are omitted. For theoretical work on the functional Taylor series expansion, we refer to von Mises (1947), Withers (1983) and Konishi (1991).

References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267–81. Budapest: Akademiai Kiado. (Reproduced (1992) in *Breakthroughs in Statistics* **1**, Ed. S. Kotz and N. L. Johnson, pp. 610–24. New York: Springer-Verlag.)

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Control* **AC-19**, 716–23.

BOZDOGAN, H. (Ed.) (1994). *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modelling: An Informational Approach*. Netherlands: Kluwer Academic.

DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73**, 323–32.

DAVISON, A. C., HINKLEY, D. V. & SCHECHTMAN, E. (1986). Efficient bootstrap simulation. *Biometrika* **73**, 555–66.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Am. Statist. Assoc.* **81**, 461–70.

GOOD, I. J. & GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–77.

GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.* **55**, 245–59.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions.* New York: Wiley.

HUBER, P. J. (1981). *Robust Statistics.* New York: Wiley.

HURVICH, C. M. & TSAI, C.-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499–509.

ISHIGURO, M., SAKAMOTO, Y. & KITAGAWA, G. (1996). Bootstrapping log-likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math.* To appear.

KITAGAWA, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with Discussion). *J. Am. Statist. Assoc.* **82**, 1032–63.

KONISHI, S. (1991). Normalizing transformations and bootstrap confidence intervals. *Ann. Statist.* **19**, 2209–25.

KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–75.

RONCHETTI, E. (1985). Robust model selection in regression. *Statist. Prob. Lett.* **3**, 21–3.

RONCHETTI, E. & STAUDTE, R. G. (1994). A robust version of Mallows's C_p. *J. Am. Statist. Assoc.* **89**, 550–9.

SAKAMOTO, Y., ISHIGURO, M. & KITAGAWA, G. (1986). *Akaike Information Criterion Statistics.* Tokyo: D. Reidel.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

SHIBATA, R. (1989). Statistical aspects of model selection. In *From Data to Model*, Ed. J. C. Willems, pp. 215–40. New York: Springer-Verlag.

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc.* B **39**, 44–7.

SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist.* A **7**, 13–26.

TAKEUCHI, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*, No. 153, pp. 12–8 (in Japanese).

TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.* **81**, 82–6.

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18**, 309–48.

WITHERS, C. S. (1983). Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. *Ann. Statist.* **11**, 577–87.

[*Received July* 1995. *Revised March* 1996]