Research Report

KSTS/RR-86/009
27 Aug., 1986

Criteria of Statistical Model Selection

by

Ritei Shibata

Department of Mathematics
Faculty of Science and Technology
Keio University

Department of Mathematics
Faculty of Science and Technology
Keio University

Hiyoshi 3-14-1, Kohoku-ku, Yokohama, 223 Japan

KSTS/RR-86/009
August 27, 1986

Criteria of Statistical Model Selection

By Ritei SHIBATA

Department of Mathematics, Keio University, Japan

ABSTRACT

Various criteria of model selection are derived from Kullback-Leibler information number. It is shown that TIC, which is an extension of Akaike's information criterion (AIC), and the cross-validation CV are asymptotically equivalent. Such criteria are further extended to RIC for the case when penalized maximum likelihood estimate is used in place of the maximum likelihood estimate. Such extension allows us to select a weight of the penalty as well as a model.

Some key words: Model Selection, AIC, TIC, RIC, CV, BIC, HQ.

1. Introduction

The objective of this paper is to draw together various criteria of model selection based on Kullback-Leibler information number and clarify the problem. Model means here a parametric family F of densities. The problem of model selection arises whenever one wants to do a statistical analysis, regression, discrimination and so on. Even the band-width selection of kernel density estimate can be considered as a kind of model selection (Rice[12]). Natural requirement for a selection procedure is to choose a model as good as possible from a given family of models. Needless to say, the goodness depends on the objective of the analysis. Two main objectives are:

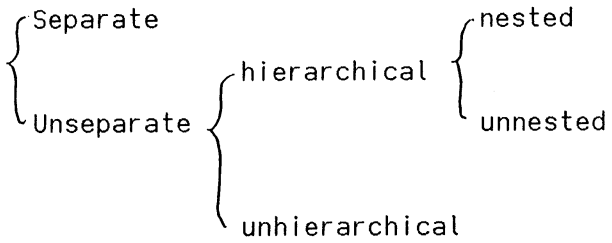
a) To know the true model $g(\cdot)$ as correctly as possible. This is what hypothesis testing aims at.

b) To choose a model which yields a good inference, for example, good estimation, good prediction and so on. The selected model is only considered as an approximation to the true $g(\cdot)$ which may not be exactly described by any of given models.

Consistency is our main concern for the former objective, and the goodness of the resulting inference is more important for the latter. However, it is not always the case that the consistency and the goodness of the inference are compatible. Recent analysis shows that some of

selection procedures are inconsistent but yield good inference for large number of observations. It will be discussed later in Section 4.

Model selection depends on the the type of family of models;



Although most of criterion procedures are formally applicable for any type of models, the difficulty of selection varies with the type of models as listed above. If models are separate, that is, each models have no intersection like normal and lognormal models, it is rather easier to distinguish models, because log of the likelihood ratio diverges to infinity with order of the number n of observations for different models (Cox[5], and Box & Cox[4]). Our theoretical interest is then mainly in unseparate models, particularly in hierarchical models.

We may choose a model by plotting values of a criterion or by looking for the minimum. An advantage of such criterion procedure is that it is easy to see differences of models by a single value. Any pre-determined values like

significance levels in multiple testings are not required. An objection to such criterion procedure may be that it obliges user to choose a unique model. However we should emphasize that the role of a criterion is not in leading analyst to a unique decision but in giving a guide of selection. The user is free to make his/her final decision by consulting any other available information as well.

In Section 2, we will derive AIC (Akaike's Information Criterion) from Kullback-Leibler information number and extend it to a criterion TIC. In Section 3, an asymptotic equivalence between the cross-validation CV and TIC will be shown. Except for independent and identically distributed observations, such equivalence does not hold true between CV and AIC. Various criteria which have been proposed will be compared in Section 4. A further extension in Section 5 allows us to select a weight of penalized maximum likelihood estimate as well as a model.

2. Information Criteria

Let $\mathbf{y}_n' = (y_1,\ldots,y_n)$ be n independent observations but not necessarily identically distributed, whose joint density is $g(\mathbf{y}_n)$, where ' denotes transpose of a vector or of a matrix. Hereafter, E denotes the expectation with respect to the vector of random variables, $\mathbf{y}_n$. Given a parametric family of densities $F = \{f(\mathbf{y}_n;\theta),\theta\in\Theta\}$, we can make use of the model F for estimation, prediction and for any other

statistical inferences. If our objective is b) in Section 1, one of natural ways of evaluating goodness of the model F is to introduce a kind of distance between the estimated density $f(\cdot;\hat{\theta})$ and the the density $g(\cdot)$, where $\hat{\theta}=\hat{\theta}(y_n)$ is the maximum likelihood estimate of θ under the model F, based on y_n. As a distance, we adopt Kullback-Leibler information number

$$K_n(g(\cdot),f(\cdot;\hat{\theta})) = \int g(x_n)\log\frac{g(x_n)}{f(x_n;\hat{\theta})}dx_n,$$

which is the function of observations y_n through $\hat{\theta}$. If

$$|\int g(x_n)\log\, g(x_n)dx_n|<\infty,$$

then the expectation of Kullback-Leibler information number $K_n(g(\cdot),f(\cdot;\hat{\theta}))$ can be rewritten as

$$\int g(x_n)\log\, g(x_n)dx_n - E\int g(x_n)\log\, f(x_n;\hat{\theta})dx_n. \qquad (2.1)$$

A statistical problem here is how to estimate (2.1), which is a function of unknown $g(\cdot)$. We will approximate (2.1) by an asymptotic expansion under the following assumptions A1 to A5.

A1. The parameter space Θ is Euclidean p-dimensional space R^p or its open subspace. Both Gradient vector

$$g_n(\theta)' = (\frac{\partial}{\partial\theta_l}l(\theta),l=1,\ldots,p)$$

and Hessian matrix

$$H_n(\theta) = \left(\frac{\partial^2}{\partial\theta_l \partial\theta_m} l(\theta), 1 \leq l, m \leq p\right)$$

of the log-likelihood function $l(\theta) = \log f(y_n;\theta)$, are well defined with probability 1, and both continuous with respect to θ.

A2. $E|g_n(\theta)| < \infty$ and $E|H_n(\theta)| < \infty$, where $|\cdot|$ denotes the absolute value of each components of a vector or of a matrix.

A3. There exists a unique θ^* in Θ, which is the solution of $Eg_n(\theta^*)=0$. This assumption together with A2 implies that $K_n(g(\cdot),f(\cdot;\theta))$ is minimized at θ^*.

A4. For any $\varepsilon>0$,

$$\sup_{\|\theta-\theta^*\|>\varepsilon} \quad l(\theta) - l(\theta^*)$$

diverges to $-\infty$ a.s.. This assumption assures that $\hat{\theta}-\theta^*$ converges to zero a.s. as n tends to infinity.

A5. For any $\varepsilon>0$, there exists $\delta>0$ such that

$$\sup_{\|\theta-\theta^*\|<\delta} \quad |E(\hat{\theta}-\theta^*)' J_n(\theta)(\hat{\theta}-\theta^*) - \mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1})| < \varepsilon$$

for large enough n. Here

$$I_n(\theta^*) = E\, g_n(\theta^*)g_n(\theta^*)' \quad \text{and} \quad J_n(\theta) = -E\, H_n(\theta)$$

are assumed positive definite matrices and continuous

with respect to θ.

We note that all of assumptions above are commonly used regularity conditions and satisfied in various situations. By expanding $\log f(x_n;\hat{\theta})$ around θ^*, we have

$$\log f(x_n;\hat{\theta}) = \log f(x_n;\theta^*) + (\hat{\theta}-\theta^*)'\frac{\partial}{\partial\theta}\log f(x_n;\theta^*)$$
$$+ \frac{1}{2}(\hat{\theta}-\theta^*)'\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x_n;\theta^{**})(\hat{\theta}-\theta^*),$$

where θ^{**} is a value between $\hat{\theta}$ and θ^*. We should note that the Gradient vector $\frac{\partial}{\partial\theta}\log f(x_n;\theta)$ and the Hessian matrix $\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x_n;\theta)$ are not of $\log f(y_n;\theta)$ but of $\log f(x_n;\theta)$. Since

$$\int g(x_n)\frac{\partial}{\partial\theta}\log f(x_n;\theta^*)dx_n = 0,$$

the assumption A3 justifies the expansion;

$$\int g(x_n)\log f(x_n;\hat{\theta})dx_n = \int g(x_n)\log f(x_n;\theta^*)dx_n$$
$$+ \frac{1}{2}(\hat{\theta}-\theta^*)'\left\{\int g(x_n)\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x_n;\theta^{**})\right\}(\hat{\theta}-\theta^*). \quad (2.2)$$

From the assumption A5, the expectation of (2.2) is

$$E\int g(x_n)\log f(x_n;\hat{\theta})dx_n = \int g(x_n)\log f(x_n;\theta^*)dx_n$$
$$- \frac{1}{2}\mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1}) + o(1)$$
$$= El(\theta^*) - \frac{1}{2}\mathrm{tr}(I_n(\theta^*)J_n(\theta^*)^{-1}) + o(1).$$

On the other hand, by expanding $l(\theta^*)$ around $\hat{\theta}$, from the fact that $\mathbf{g}_n(\hat{\theta})=0$ we have

$$l(\theta^*) = l(\hat{\theta}) + \frac{1}{2}(\theta^*-\hat{\theta})'H_n(\theta^{**})(\theta^*-\hat{\theta}). \qquad (2.3)$$

Therefore

$$E\int g(\mathbf{x}_n)\log f(\mathbf{x}_n;\hat{\theta})d\mathbf{x}_n = El(\hat{\theta}) - tr(I_n(\theta^*)J_n(\theta^*)^{-1}) + o(1),$$

and the expected Kullback-Leibler information number (2.1) becomes

$$E\ K_n(g(\cdot),f(\cdot;\hat{\theta})) = \int g(\mathbf{x}_n)\log g(\mathbf{x}_n)d\mathbf{x}_n$$

$$+ E(-l(\hat{\theta})) + tr(I_n(\theta^*)J_n(\theta^*)^{-1})) + o(1). \qquad (2.4)$$

Since the first term on the right hand side of (2.4) is independent of each model F, a natural criterion of selecting a model is derived as

$$-l(\hat{\theta}) + \overline{t_n(\theta^*)}, \qquad (2.5)$$

where $\overline{t_n(\theta^*)}$ is an estimate of $t_n(\theta^*) = tr(I_n(\theta^*)J_n(\theta^*)^{-1})$ in (2.4), which is considered as a penalty appeared in (2.2) for the increasing size of the model, plus a correction of the bias in (2.3).

There are various ways of estimating $t_n(\theta^*)$, from which different criteria may follow. If $g(\cdot)$ is equal to one of densities in F, say $f(\cdot;\theta_0)$, then $\theta^*=\theta_0$, $I_n(\theta_0)=J_n(\theta_0)$, and $t_n(\theta^*)=p$. Therefore, for the case when $g(\cdot)$ is expected

equal to or very close to one of densities in F, the criterion known as Akaike's Information Criterion (Akaike[1]),

$$AIC = -2l(\hat{\theta}) + 2p$$

follows from (2.5). Multiplication by 2 is employed only by convention.

A procedure suggested by Takeuchi[23] is to estimate $t_n(\theta^*)$ based on the sample moments, as well as based on the maximum likelihood estimate $\hat{\theta}$ of θ^*. The following example may illustrate his idea.

Example 2.1

Let us consider a simple location and scale family

$$F = \left\{ \prod_{i=1}^{n} \varphi\left[\frac{y_i-\mu}{\sigma}\right] \right\},$$

where φ is the standard normal density. In our notation, $\theta'=(\mu,\sigma)$, $\Theta = (-\infty,\infty)\times(0,\infty)$, and $f(y_i;\theta)=\varphi\left[\frac{y_i-\mu}{\sigma}\right]$. We don't assume any specific distribution for each observation but we only assume that y_i's are independent observations and have the same first and second moments. Since $\mu^* = \sum Ey_i/n$ and $\sigma^{*2} = \sum E(y_i-\mu^*)^2/n$, we have

$$\frac{1}{n} I_n(\theta^*) = \begin{bmatrix} 1/\sigma^{*2} & \mu(3)/\sigma^{*5} \\ \mu(3)/\sigma^{*5} & \mu(4)/\sigma^{*6}-1/\sigma^{*2} \end{bmatrix}$$

and

$$\frac{1}{n} J_n(\theta^*) = \begin{bmatrix} 1/\sigma^{*2} & 0 \\ 0 & 2/\sigma^{*2} \end{bmatrix},$$

where $\mu(l)=\sum_i E(y_i-\mu^*)^l/n$ for $l>1$. Then,

$$t_n(\theta^*) = 1 + \frac{1}{2}(\mu(4)/\sigma^{*4} - 1).$$

By replacing $\mu(4)$ and σ^{*2} by the 4th sample moment $\hat{\mu}(4)=\sum(y_i-\bar{y})^4/n$ and the maximum likelihood estimate $\hat{\sigma}^2=\sum(y_i-\bar{y})^2/n$ respectively, we have an estimate of $t_n(\theta^*)$,

$$\overline{t_n(\theta^*)} = 1 + \frac{1}{2}(\hat{\mu}(4)/\hat{\sigma}^4 - 1).$$

A criterion which follows from (2.5) is then

$$\begin{aligned} TIC_0(F) &= -2l(\hat{\theta}) + 2 + (\hat{\mu}(4)/\hat{\sigma}^4 - 1) \\ &= n + n\log(2\pi\hat{\sigma}^2) + 2 + (\hat{\mu}(4)/\hat{\sigma}^4 - 1). \end{aligned}$$

Multiplication by 2 is again employed by convention as is in AIC. Difference between TIC_0 and

$$AIC = -2l(\hat{\theta}) + 4$$

is clear. The effect of difference of the shape of $g(\cdot)$ from the normal is counted in TIC_0. By applying the same technique we can derive TIC_0 for the problem of selecting a sample transformation ψ. Consider models;

$$F_\psi = \left\{ \prod_{i=1}^{n} |\psi'(y_i)| \varphi\left[\frac{\psi(y_i)-\mu}{\sigma}\right] \right\},$$

where ψ' is the derivative of ψ. Then

$$TIC_0(F_\psi) = -2l(\hat{\theta}) + 2 + (\tilde{\mu}(4)/\tilde{\sigma}^4 - 1)$$

follows, where $\tilde{\mu}(4)=\sum(\psi(y_i)-\tilde{\mu})^4/n$ and $\tilde{\sigma}^2=\sum(\psi(y_i)-\tilde{\mu})^2/n$ with $\tilde{\mu}=\sum\psi(y_i)/n$. Comparing $TIC_0(F_\psi)$, we may select a transformation ψ.

However, such procedure of deriving an estimate of $t_n(\theta^*)$ is not widely applicable. It is laborious to find an estimate of $t_n(\theta^*)$ model by model, and it is usually unknown what kind of assumption is appropriate for y_i's. Before proceeding to an extension of TIC_0, we consider another example.

Example 2.2

A Gaussian regression model with m dimensional regression parameter is denoted by

$$F_m = \left\{ \prod_{i=1}^{n} \varphi\left[\frac{y_i - x_i'\beta}{\sigma}\right],\ \theta=(\beta,\sigma)' \in R^m \times (0,\infty) \right\}.$$

We first assume only independence of y_i's. Then

$$I_n(\theta^*) = \begin{bmatrix} \sum(\mu_i(2)-\mu_i(1)^2)x_i x_i'/\sigma^{*2} & \sum(\mu_i(3)-\mu_i(1)\mu_i(2))x_i'/\sigma^{*5} \\ \sum(\mu_i(3)-\mu_i(1)\mu_i(2))x_i/\sigma^{*5} & \sum(\mu_i(4)-\mu_i(2)^2)/\sigma^{*6} \end{bmatrix}$$

and

$$J_n(\theta^*) = \begin{bmatrix} X'X/\sigma^{*2} & 0 \\ 0 & 2n/\sigma^{*2} \end{bmatrix}.$$

Here $\mu_i(l) = E(e_i)^l$ for $l>1$ with $e_i=y_i-x_i'\beta^*$, $i=1, \ldots,n$ and $X=(x_1, \ldots, x_n)'$ is the design matrix.
Denoting the hat matrix by $H=(h_{ij})=X(X'X)^{-1}X'$ we have

$$t_n(\theta^*)$$

$$= \sum(\mu_i(2)-\mu_i(1)^2)h_{ii}/\sigma^{*2} + \frac{1}{2}\{\frac{1}{n}\sum(\mu_i(4)-\mu_i(2)^2)/\sigma^{*4}\}.$$

If we assume that the first and the second moments of e_i's are the same, then

$$t_n(\theta^*) = m + \frac{1}{2}(\frac{1}{n}\sum \mu_i(4)/\sigma^{*4}-1) \qquad (2.6)$$

and

$$TIC_0(F_m) = -2l(\hat{\theta}) + 2m + \frac{1}{2}(\frac{1}{n}\sum \hat{e}_i^4/\hat{\sigma}^4-1),$$

where $\hat{e}_i=y_i-x_i'\hat{\beta}$, and $\hat{\beta}$ and $\hat{\sigma}$ are the maximum likelihood estimates.
One of possible ways to avoid such specific assumption on $g(\cdot)$ is to make use of the following inequality.

$$t_n(\theta^*) \leq \sum\mu_i(2)h_{ii}/\sigma^{*2} + \frac{1}{2}(\frac{1}{n}\sum\mu_i(4)/\sigma^{*4}-1). \qquad (2.7)$$

Here the equality holds true if and only if $\mu_i(1)=E(e_i)=0$ and $\mu_i(2)=E(e_i^2)=\sigma^{*2}$ for all i, and then the value becomes to be that in (2.6). The right hand side of (2.7) can be estimated by

$$\hat{t}_n = \sum \hat{e}_i^2 h_{ii}/\hat{\sigma}^2 + \frac{1}{2}\left(\frac{1}{n}\sum \hat{e}_i^4/\hat{\sigma}^4 - 1\right).$$

This estimate is possibly biased. However it is toward safer direction. More penalty is put for models which are far from best fitting. The resulting criterion is

$$TIC(F_m) = -2l(\hat{\theta}) + 2\hat{t}_n.$$

This example leads to a general definition of TIC. Hereafter we assume that y_n is a vector of independent observations. Models should be also for independent observations, that is, the joint likelihood can be decomposed into

$$l(\theta) = \sum l_i(\theta),$$

where $l_i(\theta)=\log f_i(y_i;\theta)$. Estimate $I_n(\theta^*)$ and $J_n(\theta^*)$ by

$$\hat{I} = \sum_i \frac{\partial}{\partial\theta} l_i(\hat{\theta}) \frac{\partial}{\partial\theta'} l_i(\hat{\theta})$$

and

$$\hat{J} = -H_n(\hat{\theta}) = -\sum_i \frac{\partial^2}{\partial\theta\,\partial\theta'} l_i(\hat{\theta}),$$

respectively. Then

$$TIC = -2l(\hat{\theta}) + 2tr(\hat{I}\hat{J}^{-1})$$

is an extension of TIC_0. As noted in the example, since

$$\sum_i E\left(\frac{\partial}{\partial\theta} l_i(\theta^*) \frac{\partial}{\partial\theta'} l_i(\theta^*)\right) = I_n(\theta^*) + \sum_i E\frac{\partial}{\partial\theta} l_i(\theta^*)\; E\frac{\partial}{\partial\theta'} l_i(\theta^*)$$

(2.8)

$tr(\hat{I}\hat{J}^{-1})$ tends to over-estimate $t_n(\theta^*)$ by the last term on the right hand side of (2.8). But, if it is significant we can not expect any stable behavior of the maximum likelihood estimate $\hat{\theta}$. Each observations unevenly contribute to the Gradient of the log likelihood function at θ^*, which is the solution of

$$\sum_i E \frac{\partial}{\partial\theta} l_i(\theta^*) = 0.$$

Therefore such bias does not affect the objective to select a model which yields good inference. It is worth noting that $tr(\hat{I}\hat{J}^{-1})$ is well known Lagrange-multiplier test statistics (Hosking[8]). TIC consists of two parts, -2 log (maximum likelihood) plus twice of that test statistic.

3. Equivalence between Cross-validation and Information Criteria

Cross-validation is another kind of criterion to evaluate goodness of fit ~~of~~ a model. This criterion has a long history and has been used widely. Detailed analyses can be found in Stone[19].

We restrict our attention into a simple cross validation. By $\hat{\theta}(-i)$ we denote the maximum likelihood estimate of θ based on y_n with picking out the ith observation y_i. The cross validation is then defined as

$$CV = -2 \sum_i l_i(\hat{\theta}(-i)).$$

It is shown by Stone[20] that CV is asymptotically equivalent to AIC, when $y_1,\ldots,y_n$ are i.i.d. and $g(\cdot)$ is a member of $\mathbf{F}$. It does not hold true otherwise, but instead we can show an equivalence of CV to TIC. Necessary assumptions are the following A6 to A8 besides A1 to A4.

A6 For any $\varepsilon>0$,

$$\max_i \sup_{\|\theta-\theta^*\|>\varepsilon} (l_{-i}(\theta) - l_{-i}(\theta^*))$$

diverges to $-\infty$ a.s. as n tends to infinity, where $l_{-i}(\theta) = l(\theta)-l_i(\theta)$. This implies that $\hat{\theta}(-i)$'s, the solutions of

$$\frac{\partial}{\partial\theta} l_{-i}(\hat{\theta}(-i)) = 0 \quad i=1,\ldots,n,$$

uniformly converge to θ^* as n tends to infinity.

A7 For any $\varepsilon>0$, there exists $\delta>0$ such that

$$\sup_{\|\theta-\theta^*\|<\delta} \|I - H_n(\theta)H_n(\theta^*)^{-1}\| < \varepsilon$$

for large enough n, where $\|\cdot\|$ is Euclidean norm of a vector or the operator norm of a matrix.

A8 For any $\varepsilon>0$, there exists $\delta>0$ such that

$$\max_i \sup_{\|\theta-\theta^*\|<\delta} \|\frac{\partial^2}{\partial\theta\,\partial\theta^T} l_i(\theta)H_n(\theta^*)^{-1}\| < \varepsilon$$

for large enough n.

From the definition of $\hat{\theta}(-i)$ we have

$$\frac{\partial}{\partial\theta} l_i(\hat{\theta}(-i)) = \frac{\partial}{\partial\theta} l(\hat{\theta}(-i))$$

$$= \frac{\partial}{\partial\theta} l(\hat{\theta}) + \left\{\frac{\partial^2}{\partial\theta\partial\theta'} l(\hat{\theta})\right\}(\hat{\theta}(-i) - \hat{\theta})\ (1 + o_p(1)).$$

$$= -\hat{J}\ (\hat{\theta}(-i) - \hat{\theta})(1 + o_p(1))$$

and

$$l_i(\hat{\theta}(-i)) = l_i(\hat{\theta}) + (\hat{\theta}(-i) - \hat{\theta})' \frac{\partial}{\partial\theta} l_i(\hat{\theta})$$

$$+ (\hat{\theta}(-i) - \hat{\theta})' \frac{\partial^2}{\partial\theta\partial\theta'} l_i(\theta^{**})(\hat{\theta}(-i) - \hat{\theta})$$

$$= l_i(\hat{\theta}) - \frac{\partial}{\partial\theta'} l_i(\hat{\theta}(-i))\ \hat{J}^{-1}\ \frac{\partial}{\partial\theta} l_i(\hat{\theta})\ (1 + o_p(1))$$

$$= l_i(\hat{\theta}) - \frac{\partial}{\partial\theta'} l_i(\hat{\theta})\ \hat{J}^{-1}\ \frac{\partial}{\partial\theta} l_i(\hat{\theta})\ (1 + o_p(1)).$$

Therefore

$$CV = -2 \sum_i l_i(\hat{\theta}(-i))$$

$$= -2l(\hat{\theta}) + 2 \sum_i \frac{\partial}{\partial\theta'} l_i(\hat{\theta})\ \hat{J}^{-1}\ \frac{\partial}{\partial\theta} l_i(\hat{\theta})\ (1 + o_p(1))$$

$$= -2l(\hat{\theta}) + 2\ \mathrm{tr}(\hat{I}\hat{J}^{-1})\ (1 + o_p(1))$$

is equivalent to TIC.

Example 3.1

Consider the same regression model as in Example 2.2. To simplify our discussion, we regard σ as a nuisance parameter and estimate it by $\hat{\sigma}$. From the well known equality (Stone[19]), we have

$$CV = n \log(2\pi\hat{\sigma}^2) + \sum_i (y_i - x_i'\hat{\beta}(-i))^2/\hat{\sigma}^2$$

$$= n \log(2\pi\hat{\sigma}^2) + \sum_i \left\{\hat{e}_i/(1-h_{ii})\right\}^2/\hat{\sigma}^2.$$

To assure the consistency of $\hat{\theta}$, we assume that $\max_i(h_{ii})$ converges to zero as n tends to infinity, which is equivalent to assume A7. We have then

$$CV = n \log(2\pi\hat{\sigma}^2) + \sum_i \hat{e}_i^2(1 + 2h_{ii})/\hat{\sigma}^2 + o_p(1)$$

$$= n \log(2\pi\hat{\sigma}^2) + n + \sum_i \hat{e}_i^2 h_{ii}/\hat{\sigma}^2 + o_p(1),$$

which is asymptotically equivalent to TIC when σ is regarded as a nuisance. The term $(\frac{1}{n}\sum\hat{e}_i^4/\hat{\sigma}^4-1)$ will appear in CV, if $\hat{\sigma}(-i)$ is used in place of $\hat{\sigma}$.

GCV (Wahba[25], Li[11]) is a variant of cross validation. It is known that GCV is equivalent to AIC in the context of regression. Actually

$$GCV = \sum_i \hat{e}_i^2/(1 - m/n)^2$$

$$= \sum_i \hat{e}_i^2(1 + 2m/n) + O_p(1/n)$$

$$= n\left\{\hat{\sigma}^2(1 + 2m/n) + O_p(1/n^2)\right\},$$

and

$$AIC = n + n\log(2\pi) + 2 + n\log(\hat{\sigma}^2\exp(2m/n))$$

$$= n + n\log(2\pi) + 2 + n\log\left\{(\hat{\sigma}^2(1+2m/n)) + O_p(1/n^2)\right\}.$$

However, both criteria are different from TIC.

Although the equivalence shown above is only for the case of large enough n, this allows us more freedom to choose one of the equivalent criteria, CV or TIC. An advantage of the use of TIC is that the calculation may be simpler than that of CV. A simple reduction is possible for CV in the case of regression, but generally n maximums $l_i(\hat{\theta}(-i))$, i=1,...n should be looked for. On the other hand, only one time maximization of the likelihood is necessary for obtaining TIC. To apply TIC we have to construct beforehand a family of models which are explicitly parametrized. This seems a limitation of TIC but it should be taken as an advantage, since, in most cases, the construction of models will result in clarifying the underlying phenomena. Even in CV, the likelihood function should be explicitly parametrized. An essential limitation of TIC is that it can not be applied for the selection of models parametrized by discrete values, since differentiation of the likelihood should be allowed with respect to parameters.

4. Comparison of Criteria

To discuss the behavior of each selection procedure, it is better to classify the assumptions on the true density $g(\cdot)$.

1) $g(\cdot)$ is in a fixed family of models $\{F_j\}$ for any number n of observations.
2) $g(\cdot)$ is outside of a fixed family of models $\{F_j\}$ for an increase of n, or always stays outside.
3) $g(\cdot)$ is outside of the family $\{F_j\}$, and comes into it in the limit as the family increases its size with n.

Under the assumption 1), our first concern is about the correctness of the selection, and the goodness of the resulting inference is secondly. This assumption fits for the objective a) in Section 1. Under the assumptions 2) or 3), the correctness can not be defined well, so that our main concern is only how close the selected model is to the true $g(\cdot)$. Such assumptions fit for the objective b) in Section 1.

To discuss the consistency under the assumption 1), the following generalization of AIC (Bhansali & Downham[3], Atkinson[2], Hampel et al. [6] pp.366-367) is convenient.

$$AIC_\alpha = -2l(\hat{\theta}) + \alpha\, p,$$

where α is a pre-determined value which controls the amount

of penalty for the increasing size of the model and may depend on the size n of observations. The result by Hannan and Quinn [7] suggests that under suitable regularity conditions a necessary and sufficient condition for the strong consistency is , putting $\alpha = \alpha_n$,

$$\liminf_n \alpha_n/(2\log\log n) > 1.$$

and

$$\limsup_n \alpha_n/n = 0.$$

That for the weak consistency is

$$\liminf_n \alpha_n = \infty$$

and

$$\limsup_n \alpha_n/n = 0.$$

The result above is not yet generally proved, but intuitively clear if we note that $2\{l(\hat{\theta})-l(\theta_0)\}$ is χ^2 distributed with the degree of freedom p if $g(\cdot)$ is equal to $f(\cdot;\theta_0)$, a density in the underlying model, otherwise χ^2 distributed with a degree of freedom with the order of n. The condition for strong consistency comes from the law of iterated logarithm. Therefore, the AIC, TIC or CV introduced in the previous section are not consistent. For the asymptotic distribution, see Shibata [14], Bhansali & Downham[3], and Woodroofe[26]. They obtained the asymptotic distribution of

the selected model by applying theorems of random walk. Some of consistent criterion procedures have been proposed, BIC by Schwarz[13] and HQ by Hannan and Quinn[7], which are AIC_α with $\alpha = \log n$ and $\alpha = c \log\log n$ for $c > 2$, respectively. It is interesting note the result by Takada[22], that any procedure so as to minimize AIC_α is admissible under the 0-1 loss. In other word, if our main concern is the correctness of the selection, there is no dominant selection procedure in such class of selection procedures.

If we put assumption 2), at least asymptotically, the largest model or the full model which includes any underlying models, will be the best possible selection. Apparently such full model can be constructed from the given family of models.

Under the assumption 3), our main concern is about the goodness of the resulting inference rather than the correctness. The key point for proving an optimality property of AIC is that the trade off between the bias and the variance remains significant even when n is large enough. If we restrict our attention into the estimation of regression parameters, such trade off mechanism is rigorously formulated. The result by Shibata[15] shows that if the regression variables are selected so as to minimize one of AIC_α then the selection is asymptotically optimal if and only if $\alpha=2$, that is the case of AIC. Necessary assumptions for the proof are that the shape of $g(\cdot)$ is the same as that of F, and the

mean vector of observations is parametrized by infinitely many regression parameters. Otherwise, AIC is not necessarily optimal. But TIC is instead expected optimal under the loss function like, Kullback-Leibler information number as well as under the squared loss, even when the shape of $g(\cdot)$ does not coincide with that in $\mathbf{F}$. This results will be reported elsewhere.

For admissibility under the squared loss with an additional penalty p, Stone[21] proved local asymptotic admissibility, and Kempthorne[10] proved the admissibility under the squared loss. Such results are corresponding to the result by Takada[22] in the case of 0-1 loss function.

All of the results above is in the sense of asymptotics. If the size n is fixed, theoretical comparison is difficult and only available results are by simulations. Recent paper by Hurvich[9] will help the understanding of the behavior in small samples, for example, consistency does not necessarily imply the goodness of selection. One of practical procedure might be obtained by choosing α according to the size n (see Shibata[16]).

For more detailed discussion on incompatibility between consistency and efficiency, see Shibata[18], and for comparisons with testing procedures see Shibata[17].

5. Further extension

In previous sections, the maximum likelihood estimate has been always used as an estimate of parameter θ. In fact, there is no definite reason why we restrict our attention into such estimate, though, the maximum likelihood estimate has nice properties and is convenient for analysis. As far as $f(\cdot;\hat{\theta})$ is close to $g(\cdot)$, we may use any other estimate in place of $\hat{\theta}$ under each model. In this section, we consider the penalized maximum likelihood estimate and derive a criterion based on such estimate. A typical example of penalized estimate is ridge regression or spline function approximation. A crucial problem in such estimate is how to choose a weight of penalty. As a result of the extension, we can unify selection of such weight and that of a model.

Penalized likelihood is defined as

$$L_\lambda(y_n;\theta) = \log f(y_n;\theta) + \lambda k(\theta),$$

where $k(\theta)\leqq 0$ is an arbitrary penalty function which may depend on n and twice differentiable. The weight $\lambda\geqq 0$ controls the amount of penalty.

The penalized maximum likelihood estimate $\hat{\theta}(\lambda)$ is the solution of

$$\frac{\partial}{\partial\theta}L_\lambda(y_n;\theta) = 0.$$

We assume that $\hat{\theta}(\lambda)$ converges to $\theta^*(\lambda)$ which is the unique solution of

$$E \frac{\partial}{\partial\theta} L_\lambda(y_n;\theta) = 0.$$

By similar expansions as in Section 2, we can show

$$E \int L_\lambda(x_n;\hat{\theta}(\lambda))g(x_n)dx_n$$

$$= E\, L_\lambda(y_n;\hat{\theta}(\lambda)) - E\, (\hat{\theta}(\lambda)-\theta^*(\lambda))' J_n(\lambda)(\hat{\theta}(\lambda)-\theta^*(\lambda)) + o(1), \tag{5.1}$$

where

$$J_n(\lambda) = - E \frac{\partial^2}{\partial\theta\partial\theta'} L_\lambda(y_n;\theta^*(\lambda)).$$

Subtracting $\lambda k(\hat{\theta}(\lambda))$ from the both sides of (5.1), we have

$$E \int g(x_n)\log f(x_n;\hat{\theta}(\lambda))dx_n$$

$$= E\{ l(\hat{\theta}(\lambda)) - (\hat{\theta}(\lambda)-\theta^*(\lambda))' J_n(\lambda)(\hat{\theta}(\lambda)-\theta^*(\lambda)) \} + o(1).$$

Since the expansion

$$0 = \frac{\partial}{\partial\theta} L_\lambda(y_n;\theta^*(\lambda)) + \frac{\partial^2}{\partial\theta\partial\theta'} L_\lambda(y_n;\theta^*(\lambda))(\hat{\theta}(\lambda)-\theta^*(\lambda))$$

asymptotically holds true, we can rewrite the expectation of Kullback-Leibler information number as

$$E \int \log \frac{g(x_n)}{f(x_n;\hat{\theta}(\lambda))}\, g(x_n)dx_n$$

$$= \int g(x_n)\log g(x_n)dx_n - E\, l(\hat{\theta}(\lambda)) + \mathrm{tr}(I_n(\lambda)J_n(\lambda)^{-1}) + o(1),$$

where

$$I_n(\lambda) = E \frac{\partial}{\partial\theta} L_\lambda(y_n;\theta^*(\lambda)) \frac{\partial}{\partial\theta'} L_\lambda(y_n;\theta^*(\lambda)).$$

Then an extension of TIC follows,

$$RIC = -2l(\hat{\theta}(\lambda)) + 2tr(\hat{I}(\lambda)\hat{J}(\lambda)^{-1}),$$

where

$$\hat{I}(\lambda) = \sum\{\frac{\partial}{\partial\theta} l_i(\hat{\theta}(\lambda)) + \frac{\lambda}{n} \frac{\partial}{\partial\theta} k(\hat{\theta}(\lambda))\}\{\frac{\partial}{\partial\theta} l_i(\hat{\theta}(\lambda)) + \frac{\lambda}{n} \frac{\partial}{\partial\theta} k(\hat{\theta}(\lambda))\}'$$

$$= \sum \frac{\partial}{\partial\theta} l_i(\hat{\theta}(\lambda)) \frac{\partial}{\partial\theta'} l_i(\hat{\theta}(\lambda)) - \frac{\lambda^2}{n} \frac{\partial}{\partial\theta} k(\hat{\theta}(\lambda)) \frac{\partial}{\partial\theta'} k(\hat{\theta}(\lambda))$$

and

$$\hat{J}(\lambda) = -\frac{\partial^2}{\partial\theta\partial\theta'} l(\hat{\theta}(\lambda)) - \lambda \frac{\partial^2}{\partial\theta\partial\theta'} k(\hat{\theta}(\lambda)).$$

When λ=0, RIC is reduced to TIC, so that RIC is in fact an extension of TIC. By RIC, we can choose λ, as well as selecting a model. One of practical procedures is to choose λ for each model so as to minimize RIC and compare the minimized value of RIC for each model. There is no rigorous proof of optimality yet, but it is clear from the derivation that we are looking for an $f(\cdot;\theta)$ in given models $\{F_j\}$, as close as possible to $g(\cdot)$ in terms of Kullback-Leibler information number.

Example 5.1

Consider the same regression model as in Example 2.2. To simplify the problem, we regard σ as a nuisance parameter. As a penalty function we adopt

$$k(\theta) = - \|X\beta\|^2/2\sigma^2 .$$

The penalized maximum likelihood estimate of β is then a shrinkage estimate, $\hat{\beta}(\lambda)=\hat{\beta}(0)/(1+\lambda)$, where $\hat{\beta}(0)$ denotes the maximum likelihood estimate of β. Since

$$\hat{I}(\lambda) = \sum \hat{e}_i^2 \, x_i x_i'/\sigma^4$$

and

$$\hat{J}(\lambda) = (1+\lambda) \, X'X/\sigma^2 ,$$

we have

$$RIC(F_m,\lambda) = n \log 2\pi\sigma^2 + \Sigma(y_i - x_i'\hat{\beta}(\lambda))^2/\sigma^2 + \frac{2}{1+\lambda}\Sigma\hat{e}_i^2 h_{ii}/\sigma^2$$

$$= n \log 2\pi\sigma^2 + \{ \Sigma\hat{e}_i^2 + (\frac{\lambda}{1+\lambda})^2\Sigma\hat{y}_i^2 + \frac{2}{1+\lambda} \Sigma\hat{e}_i^2 h_{ii} \}/\sigma^2 ,$$

where $\hat{y}_i = y_i - \hat{e}_i$. Here

$$\frac{\partial}{\partial\theta} RIC(F_m,\lambda) = \frac{1}{(1+\lambda)^3\sigma^2}\{ \lambda(\Sigma\hat{y}_i^2 - \Sigma\hat{e}_i^2 h_{ii}) - \Sigma\hat{e}_i^2 h_{ii} \} \quad (5.2)$$

The $\hat{\lambda}$ which minimizes RIC is then

$$\hat{\lambda} = \frac{\Sigma\hat{e}_i^2 h_{ii}}{\Sigma\hat{y}_i^2 - \Sigma\hat{e}_i^2 h_{ii}} \qquad \text{if} \qquad \Sigma\hat{y}_i^2 > \Sigma\hat{e}_i^2 h_{ii},$$

$$= \infty \qquad \text{otherwise.}$$

The resulting estimate of β is

$$\hat{\beta}(\hat{\lambda}) = (1 - \Sigma\hat{e}_i^2 h_{ii}/\Sigma\hat{y}_i^2)^+ \hat{\beta}(0),$$

where $(\alpha)^+ = \max(\alpha,0)$. It is interesting to note that a non-negative shrinkage factor automatically follows from

minimizing RIC. As a special case, for the model with a single location parameter μ as in Example 2.1,

$$\hat{\mu}(\hat{\lambda}) = (1 - \hat{\sigma}^2/n\bar{y}^2)^+ \bar{y},$$

which is one of commonly used shrinkage estimates.
The minimum value of RIC for each model is

$$RIC(F_m,\hat{\lambda}) = n \log 2\pi\sigma^2 + \{ \sum\hat{e}_i^2 + 2(1 - \frac{1}{2}\frac{\sum\hat{e}_i^2 h_{ii}}{\sum\hat{y}_i^2})(\sum\hat{e}_i^2 h_{ii}) \}/\sigma^2$$

$$\text{if } \sum\hat{y}_i^2 > \sum\hat{e}_i^2 h_{ii} \quad \text{i.e. } \hat{\lambda} < \infty,$$

$$= n \log 2\pi\sigma^2 + \sum y_i^2/\sigma^2 \quad \text{otherwise.}$$

Thus

$$RIC(F_m,\infty) \leqq RIC(F_m,\hat{\lambda}) \leqq RIC(F_m,0),$$

and $RIC(F_m,\lambda)$ decreases as λ increases from 0 and attains the minimum at $\hat{\lambda}$. Particularly when $\hat{\lambda}=\infty$, the complete shrinkage estimate $\hat{\beta}(\infty)=0$ follows. By using such estimate we can always decrease the value of RIC except for the case when all $\hat{e}_i$'s are zero. We then compare such minimized value for different models F_m, and choose one of them.
More generally if the penalty function is of the form of $k(\theta) = -\|A\beta\|^2/2\sigma^2$, then

$$RIC(F_m,\lambda) = n \log 2\pi\sigma^2 + \{ \|y - X\hat{\beta}(\lambda)\|^2 + 2\sum h_{ii}(\lambda)\hat{e}_i^2 \}/\sigma^2,$$

where

$$H(\lambda) = (h_{ij}(\lambda)) = X(X'X+\lambda A'A)^{-1}X'$$

and

$$\hat{\beta}(\lambda) = (X'X+\lambda A'A)^{-1}X'y.$$

As a result, in this regression context, RIC is closely related to a criterion $\hat{T}(h)$ which is mentioned in Titterington[24].

It is also possible to extend RIC for the case of more than one penalty functions. Still much works should be done for this criterion. We leave those for future investigations.

References

[1] Akaike, H., "Information theory and an extension of the maximum likelihood principle," pp. 267-281 in 2nd Int. Symposium on Information Theory, ed. Petrov B.N. and Csáki, F., Akadémia Kiado, Budapest (1973).

[2] Atkinson, A. C., "A note on the generalized information criterion for choice of a model," Biometrika Vol. 67, pp. 413-418 (1980).

[3] Bhansali, R. J. and D. Y. Downham, "Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion," Biometrika Vol. 64, pp. 547-551 (1977).

[4] Box, G. E. P. and D. R. Cox, "An analysis of transformations," J. Roy. Statist. Soc., pp. 211-243 (1964).

[5] Cox, D. R., "Further results on tests of separate families of hypotheses," J. Roy. Statist. Soc., pp. 406-424 (1962).

[6] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel, Robust Statistics: the approach based on influence functions, John Wiley (1986).

[7] Hannan, E. J. and B. G. Quinn, "The determination of the order of an autoregression," J. Roy. Statist. Soc. Vol. **B 41**, pp. 190-195 (1979).

[8] Hosking, J. R. M., "Lagrange-multiplier tests of time-series models," J. R. Statist. Soc. Vol. **B42**, pp. 170-181 (1980).

[9] Hurvich, C. M., "Data-Driven choice of a spectraum estimate: Extending the applicability of cross-validation methods," J. Amer. Statist. Soc. Vol. **80**, pp. 933-940 (1985).

[10] Kempthorne, P. J., "Admissible variable-selection procedures when fitting regression models by least squares for prediction," Biometrika Vol. **71**, pp. 593-597 (1984).

[11] Li, K., "From Stein's unbiased estimates to the method of generalized cross validation," Ann. Statist. Vol. **13**, pp. 1352-1377 (1985).

[12] Rice, J., "Bandwidth choice for nonparametric kernel regression," Ann. Statist. Vol. **12**, pp. 1215-1230 (1984).

[13] Schwarz, G., "Estimating the dimension of a model," Ann. Statist. Vol. **6**, pp. 461-464 (1978).

[14] Shibata, R., "Selection of the order of an autoreegressive model by Akaike's information criterion," Biometrika Vol. **63**, pp. 117-126 (1976).

[15] Shibata, R., "An optimal selection of regression variables," Biometrika Vol. **68**, pp. 45-54, Correction 69, p.492 (1981).

[16] Shibata, R., "Selection of the number of regression variables; a minimax choice of generalized FPE," to appear in Ann. Inst. Statist. (1985).

[17] Shibata, R., "Selection of regression variables," in Encyclopedia of Statistical Sciences, John Wiley & Sons (1986). (to appear)

[18] Shibata, R., "Consistency of model selection and that of parameter estimateion," pp. 127-141 in Essays in Time Series and Allied Processes, ed. J. M. Gani and M. B. Priestley, Applied Probability Trust, Sheffield

(1986).

[19] Stone, M., "Cross-validatory choice and assessment of statistical predictions," J. Roy. Statist. Soc. Vol. **36**, pp. 111-133 (1974).

[20] Stone, M., "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," J. Roy. Statist. Soc. Vol. **B 39**, pp. 44-47 (1977).

[21] Stone, C. J., "Local asymptotic admissibility of a generalization of Akaike's model selection rule," Ann. Inst. Statist. Math. Vol. **34**, pp. 123-133 (1982).

[22] Takada, Y., "Admissibility of some variable selection rules in linear regression model," J. Japan Statist. Soc. Vol. **12**, pp. 45-49 (1982).

[23] Takeuchi, K., "Distribution of information statistics and a criterion of model fitting," Suri-Kagaku (Mathematical Sciences) Vol. **153**, pp. 12-18, (in Japanese) (1976).

[24] Titterington, D. M., "Common structure of smoothing techniques in statistics," Int. Statist. Rev. Vol. **53**, pp. 141-170 (1985).

[25] Wahba, G., "A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem," Ann. Statist. Vol. **13**, pp. 1378-1402 (1985).

[26] Woodroofe, M., "On model selection and the arc sine laws," Ann. Statist. Vol. 10(4), pp. 1182-1194 (1982).