BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-88/21

COMPARING NOT NECESSARILY NESTED MODELS WITH THE MINIMUM AIC AND THE MAXIMUM KULLBACK-LEIBLER ENTROPY CRITERIA: NEW PROPERTIES AND CONNECTIONS

by

David F. Findley
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Recommended by: Nash J. Monsour

Report completed: September 6, 1988

Report issued: September 6, 1988

COMPARING NOT NECESSARILY NESTED MODELS WITH THE MINIMUM AIC AND THE MAXIMUM KULLBACK–LEIBLER ENTROPY CRITERIA: NEW PROPERTIES AND CONNECTIONS

David F. Findley

ABSTRACT

Applied statistical modelers frequently have to compare models of rather different forms. To the extent that objective criteria are used to facilitate such comparisons, Akaike's minimum AIC criterion seems to be the one most widely used, due in part, perhaps, to its ease of use and its impressive successes in some industrial applications. A coherent theory to motivate MAIC's use with non–nested model comparisons has been lacking, however, and the present paper seeks to describe one. Not surprisingly, Akaike's non–operative Entropy Maximization Principle turns out to provide a model of what successful performance might mean in some subtle situations involving incorrect models. This paper summarizes some new results concerning this principle, a linear stochastic regression version of Akaike's criterion, and the related criteria of Schwarz and Hannan and Quinn. Some analyses related to a successful ship autopilot design project are presented to illustrate the application of MAIC. Our theoretical results are directed towards analyzing the performance of the model selection criteria in some general situations, including three in which the preferred model, or lack of one, seems obvious a priori. Loosely described, these three situations are:

(i) The best model from one class fits the data better than all models from the other class.

(ii) Both model classes include the correct model, but one class has fewer parameters to be estimated than the other.

(iii) The two model classes have the same number of parameters to be estimated and both include the correct model.

We also analyze the performance of the various criteria in two situations in which the principle of parsimony is contradicted.

1. Introduction

We assume that a q–dimensional process y_t is given which is to be linearly regressed upon one or more other processes (stochastic, nonstochastic or mixed). If the r–dimensional process x_t is a candidate regressor, then several points of view are possible concerning the "ideal" regression relation to be estimated,

$$(1.1) \qquad y_t = A^{(x)}x_t + e_t^{(x)} ,$$

beyond the basic assumptions that the error process $e_t^{(x)}$ has mean zero and a constant, nonsingular covariance matrix $\Sigma^{(x)}$.

One point of view is that (1.1) simply represents a potentially useful approximation formula, in which case the ideal regression function $A^{(x)}x_t$ need not be subjected to very demanding conditions: some useful results can be obtained from the simple uncorrelatedness requirement

$$(1.2) \qquad Ee_t^{(x)}x_t' = 0 \ .$$

At the other extreme, we may wish $A^{(x)}x_t$ to capture information about y_t in a quite strong sense. In this case, motivated by stochastic control problems, we will assume that the entries of x_t are variables which were measured or calculated at time t–1 or before, and we will let I_t denote the information set (σ–algebra) generated by <u>all</u> such variables, <u>including those not in</u> x_t <u>which might be used as components in some other competing regressor</u> $\tilde{x}_t$. We will say that x_t contains the correct regression variables, or is a <u>complete regressor</u>, if the mean of y_t conditional on the information set I_t is a linear function of the entries of x_t, so that, for some (qxr)–matrix $A^{(x)}$,

$$(1.3) \qquad A^{(x)}x_t = E(y_t|I_t) \quad \text{(for all t)},$$

holds, and if, also, the variability of $e_t^{(x)}$ is unaffected by I_t, in the sense that

$$(1.4) \qquad E(e_t^{(x)}e_t^{(x)\prime}|I_t) = \Sigma^{(x)} \text{ (for all t)}.$$

If, in addition, all columns of $A^{(x)}$ are non–zero, then x_t will be called a <u>correct regressor</u>.

All complete regressors are <u>equivalent</u> regressors for y_t in the (asymptotic) sense that they have the same values of $A^{(x)}x_t$ and $e_t^{(x)}$. The conditions (1.3) and (1.4) will be satisfied, for example, if the error $e_t^{(x)}$ is independent of the variates generating I_t.

We will be concerned with least squares estimates of $A^{(x)}$ and $\Sigma^{(x)}$ from data x_t, y_t, $t=1,\ldots,N$. These are the matrices $\hat{A}_N$ and $\hat{\Sigma}_N$ which maximize the Gaussian log–likelihood function[1]

$$\begin{aligned} L_N[\Sigma,A] &= -(N/2)\log 2\pi|\Sigma| \\ &\quad - (1/2)\sum_{t=1}^{N}(y_t - Ax_t)'\Sigma^{-1}(y_t - Ax_t) \ , \end{aligned}$$

and which are given by

$$\hat{A}_N = (\sum_{t=1}^{N} y_t x_t')(\sum_{t=1}^{N} x_t x_t')^{-1}$$

and

$$\hat{\Sigma}_N = N^{-1}\sum_{t=1}^{N}(y_t - \hat{A}_N x_t)(y_t - \hat{A}_N x_t)' \ .$$

The maximized log–likelihood is

$$L_N[\hat{\Sigma}_N,\hat{A}_N] = -(N/2)\{\log 2\pi|\hat{\Sigma}_N| + q\} \ .$$

From (1.1) and (1.2) and the assumption that $Ex_t x_t'$ is nonsingular, we obtain

[1] This is an extension of the usual terminology, since $L_N[\Sigma,A]$ is not a log joint density function for the observed values of x_t and y_t except in the special case in which x_t is deterministic. It is a sum of N terms which can be interpreted as log Gaussian conditional densities for the individual y_t's.

$$A^{(x)} = (Ey_t x_t')(Ex_t x_t')^{-1} \quad . \tag{1.5}$$

Note that if x_t and y_t are jointly covariance stationary (with mean 0) and if we define $A^{(x)}$ as in (1.5) and set $e_t^{(x)} = y_t - A^{(x)}x_t$, then (1.1) and (1.2) hold. For example, if y_t is a scalar, mean–zero, stationary process and $x_t = y_{t-k}$, then $A^{(x)} = \rho_k$, the autocorrelation at lag k.

If y_t is nonstationary with changing mean or variance, then x_t must capture this effect well enough that the error process $e_t^{(x)}$ satisfies our zero mean and constant variance requirements. For example, if y_t is a nonstationary scalar process whose increments $y_t - y_{t-1}$ are stationary, then x_t should have y_{t-1} as one of its components, see Tiao and Tsay (1983).

Under (1.2), the matrix pair $\Sigma^{(x)}$, $A^{(x)}$ can be shown to be the unique maximizer of the version of the <u>Kullback–Leibler entropy functional</u> defined by taking the expected value of the log–likelihood function with respect to the true distribution of the data:

$$E_N[\Sigma,A] = E\{L_N[\Sigma,A]\} = -(N/2)(\log 2\pi|\Sigma|$$

$$+\mathrm{tr}\Sigma^{-1}E\{N^{-1}\sum_{t=1}^{N}(y_t - Ax_t)(y_t - Ax_t)'\})$$

(tr denotes trace). Observe that

$$E_N[\Sigma^{(x)},A^{(x)}] = -\ (N/2)\ \{\log 2\pi|\Sigma^{(x)}| + q\} \ .$$

Even with incomplete regressors, $\hat{\Sigma}_N$ and $\hat{A}_N$ are consistent estimators of $\Sigma^{(x)}$ and $A^{(x)}$ under quite general circumstances, that is, $\hat{\Sigma}_N \rightarrow_p \Sigma^{(x)}$ and $\hat{A}_N \rightarrow_p A^{(x)}$, see Hannan (1970), Hosoya and Taniguchi (1982), Lai and Wei (1982) and Ljung and Caines (1979). Therefore the entropy functional $E_N[\Sigma,A]$ can be said to identify the asymptotic values of the maximum likelihood estimates.

Our goal is to derive some properties of entropy and log–likelihood which are relevant for model selection, particularly regressor selection. Given two candidate regressor processes $x_t^{(i)}$, i=1,2, let us denote by $\hat{\Sigma}_N^{(i)}$, $\Sigma^{(i)}$, $\hat{A}_N^{(i)}$, $A^{(i)}$, i=1,2, their corresponding values of $\hat{\Sigma}_N$, $\Sigma^{(x)}$, $\hat{A}_N$ and $A^{(x)}$. An elementary argument can be used to show that if $x_t^{(1)}$ is a complete regressor but $x_t^{(2)}$ is not, then $|\Sigma^{(1)}|<|\Sigma^{(2)}|$, which is equivalent to $E_N[\Sigma_N^{(1)},A_N^{(1)}] > E_N[\Sigma_N^{(2)},A_N^{(2)}]$. Thus, larger entropy values are associated with the preferred regressor. Of course, there are further considerations: a correct regressor would be preferred over a regressor which was complete but not correct (and therefore contains superfluous variables) although both would have the same value of $\Sigma^{(x)}$; also, $\Sigma^{(i)}$ and $A^{(i)}$ are not known, only estimates $\hat{\Sigma}_N^{(i)}$, $\hat{A}_N^{(i)}$ i=1,2 are available, etc. Akaike's *entropy maximization principle* (EMP) asserts that, among estimated models, the one with largest entropy is to be preferred. For regressor selection, this becomes

(EMP). *Prefer* $x_t^{(1)}$ *over* $x_t^{(2)}$ *if the entropy difference*

$$E_N[\hat{\Sigma}_N^{(1)},\hat{A}_N^{(1)}] - E_N[\hat{\Sigma}_N^{(2)},\hat{A}_N^{(2)}] \tag{1.6}$$

is positive.

This is not an operative criterion, because the entropy difference cannot be calculated, the true distribution of the observed processes being unknown. The usual interpretation of Akaike's work (1973) is that *the sign of* (1.6) *can be estimated by the sign of*

$$L_N[\hat{\Sigma}_N^{(1)},\hat{A}_N^{(1)}] - L_N[\hat{\Sigma}_N^{(2)},\hat{A}_N^{(2)}] - q(r^{(1)} - r^{(2)})$$

(1.7)

$$= -\ (N/2)\ \log(\hat{\Sigma}_N^{(1)}|/|\hat{\Sigma}_N^{(2)}|) - q(r^{(1)} - r^{(2)})$$

where $r^{(i)} = \dim x_t^{(i)}$, $i=1,2$.

To describe the connection between (1.7) and our version of Akaike's Minimum AIC criterion, we require the fourth cumulants of $\tilde{e}_t = (\Sigma^{(x)})^{-1/2}e_t^{(x)}$. (We use $\Sigma^{1/2}$ to designate a matrix with the property that $\Sigma = (\Sigma^{1/2})(\Sigma^{1/2})'$ and $\Sigma^{-1/2}$ to denote its inverse.) With $\tilde{e}_{jt}$ denoting the j–th entry of $\tilde{e}_t$, $1 \le j \le q$, these are given by

$$\kappa_{jk} = E(\tilde{e}_{jt}^2 \tilde{e}_{kt}^2) - 1 - 2\delta_{jk}\ ,$$

where δ_{jk} is 1 if j=k and 0 otherwise. Then, for the purpose of theoretical analysis, we define

(1.8)

$$\begin{aligned} AIC_N^{(i)} &= -2L_N[\hat{\Sigma}_N^{(i)},\hat{A}_N^{(i)}] \\ &+ 2\{q(q+1)/2 + qr^{(i)}\} + \sum_{j,k=1}^{q} \tilde{\kappa}_{jk}^{(i)} \end{aligned}$$

(the term in curly braces is the number of distinct parameters estimated in $\hat{\Sigma}_N^{(i)}$ and $\hat{A}_N^{(i)}$, i=1,2), and set

$$\mathrm{DAIC}_N^{(1,2)} = \mathrm{AIC}_N^{(1)} - \mathrm{AIC}_N^{(2)}.$$

The fourth–cumulant sum in (1.8), which does not appear in the definition of AIC in the traditional correct–model maximum likelihood estimation context, see Akaike (1973) for example, occurs here because a Gaussian likelihood function has been used to model possibly non–Gaussian data. If the error processes $e_t^{(1)}$ and $e_t^{(2)}$ are Gaussian, the fourth cumulant expressions are zero and (1.7) coincides with $(-1/2)\mathrm{DAIC}_N^{(1,2)}$. If the processes $x_t^{(1)}$ and $x_t^{(2)}$ are complete, they are <u>equivalent</u> in the sense that $A^{(1)}x_t^{(1)} = A^{(2)}x_t^{(2)}$ and (therefore) $e_t^{(1)} = e_t^{(2)}$ hold (almost surely) for all t. Thus, for equivalent regressor processes too, $(-1/2)\mathrm{DAIC}_N^{(1,2)}$ coincides with (1.7). Nearly equivalent regressor can be expected to have this property to a good approximation. Our purpose in including the cumulant expression in the definition of AIC will become apparent in section 3b below, where it will be seen that modeling contraints imposed on the error covariance matrix, such as (3.2), interfere with the properties of $\mathrm{DAIC}_N^{(1,2)}$ unless certain cumulants are 0, see (3.4). The <u>minimum AIC criterion for regressor selection</u> is:

(MAIC). <u>Prefer the regressor</u> $x_t^{(1)}$ <u>over</u> $x_t^{(2)}$ <u>if</u> $\mathrm{DAIC}_N^{(1,2)}$ <u>is negative</u>.

2. Overview of the Main Results of (4)

Let us abbreviate $L_N[\hat{\Sigma}_N,\hat{A}_N]$, $E_N[\hat{\Sigma}_N,\hat{A}_N]$, $L_N[\Sigma^{(x)},A^{(x)}]$ and $E_N[\Sigma^{(x)},A^{(x)}]$ by $\hat{L}_N$, $\hat{E}_N$, $L_{N,\infty}$ and $E_{N,\infty}$, respectively, adding a superscript $^{(i)}$ when the regressor $x_t^{(i)}$ is considered ($\hat{L}_N^{(i)}$, etc.). All of our analyses will be connected with the terms of the fundamental decompositions (LD) and (ED):

$$\text{(LD)} \qquad \hat{L}_N = \{\hat{L}_N - L_{N,\infty}\} + \{L_{N,\infty} - E_{N,\infty}\} + E_{N,\infty} \ ,$$

$$\text{(ED)} \qquad \hat{E}_N = \{\hat{E}_N - E_{N,\infty}\} + E_{N,\infty} \ .$$

The term $E_{N,\infty}$, being proportional to N, is the dominant term in these decompositions. (In familiar situations, the fundamental log–likelihood ratio $\hat{L}_N - L_{N,\infty}$ converges in distribution (to χ^2_{qr}, usually) and so is bounded in probability, and the same is true of $\hat{E}_N - E_{N,\infty}$, whereas the term $L_{N,\infty} - E_{N,\infty}$ is of order $N^{1/2}$ in probability.) Thus, when $|\Sigma^{(1)}| \neq |\Sigma^{(2)}|$, $E^{(1)}_{N,\infty} - E^{(2)}_{N,\infty}$ dominates the behaviors of both $\hat{E}^{(1)}_N - \hat{E}^{(2)}_N$ and $DAIC^{(1,2)}_N$, see (2.3) below.

The situation with equivalent regressors is more subtle. If $x^{(1)}_t$ and $x^{(2)}_t$ are equivalent, then $E^{(1)}_{N,\infty} = E^{(2)}_{N,\infty}$ and $L^{(1)}_{N,\infty} = L^{(2)}_{N,\infty}$, so that the behaviors of $\hat{E}^{(1)}_N - \hat{E}^{(2)}_N$ and $DAIC^{(1,2)}_N$ depend only on properties of the first terms on the right in (LD) and (ED).

The crucial fact is that these first terms behave oppositely. This is always true in the simple sense that $\hat{L}_N - L_{N,\infty}$ is positive (since $\hat{\Sigma}_N$ $\hat{\Sigma}_N$, $\hat{A}_N$ maximizes $L_N[\Sigma, A]$) whereas $\hat{E}_N - E_{N,\infty}$ is negative ($\Sigma^{(x)}$, $A^{(x)}$ maximizes $E_N[\Sigma,A]$). However, in many, although not all, important situations, a deeper result holds:

$$\text{(A–S)} \qquad \{\hat{L}_N - L_{N,\infty}\} + \{\hat{E}_N - E_{N,\infty}\} \rightarrow_{p,E} 0 \ .$$

Here $\rightarrow_{p,E}$ denotes convergence both in probability and in expectation. (This seems to have been recognized first by Akaike, and first rigorously proved, for the case of complete univariate autoregressors in the Gaussian context, by R. Shimizu (1978), who discussed only convergence in probability.) Observe that if $x^{(1)}_t$ and $x^{(2)}_t$ are equivalent regressors for which (A–S) holds, then subtracting the two (A–S) expression yields the fact that the log–likelihood ratio $\hat{L}_N^{(1)} - \hat{L}_N^{(2)}$ and the entropy difference behave oppositely in the strong sense that

$$\{\hat{E}_N^{(1)} - \hat{E}_N^{(2)}\} - \{\hat{L}_N^{(2)} - \hat{L}_N^{(1)}\} \to_p 0 \, . \tag{2.1}$$

This "opposite" behavior will be seen to enable (EMP) to make correct decisions in situations in which superfluous variables are present. (Interestingly, (2.1) also holds for an important class of nonstationary regressors for which (A–S) fails, this being the class of unstable autoregressions studied by Chan and Wei (1988).)

2a. Comparing Regressors: To show the power of (EMP) we will now list the situations in which most statistical researchers would be willing to make decisions *a priori* about regressors. We are assuming that no special considerations apply, such as one regressor's being significantly more difficult or expensive to obtain than the other. We will, at the same time, describe senses in which (EMP) and, to a somewhat lesser extent, (MAIC) lead to correct decisions in these situations. These are the main results of the paper. The precise hypotheses required for the results are given in Findley and Wei (1988).

First we consider the neutral situation, in which neither regression would be preferred over the other.

(AP0). *The processes* $x_t^{(1)}$ *and* $x_t^{(2)}$ *are complete and have the same dimension,* $r^{(1)} = r^{(2)}$.

In this situation, it can be shown that, asymptotically, (EMP) _and_ (MAIC) _have no preference for one regressor over the other_. More precisely, there is a random variable S with symmetric distribution, $\Pr\{S>\alpha\} = P\{S<-\alpha\}$ for all $\alpha\geq 0$, such that

$$(2.2) \qquad (1/2)\mathrm{DAIC}_N^{(1,2)},\ \hat{E}_N^{(1)} - \hat{E}_N^{(2)} \rightarrow_{\mathrm{dist.}} S\ .$$

Thus, in particular, for large enough N, a realization of the processes y_t, $x_t^{(1)}$, $x_t^{(2)}$, $t=1,\dots,N$ for which $x_t^{(1)}$ is preferred, by (EMP) or (MAIC), is approximately as probable as one for which $x_t^{(2)}$ is preferred. (S has a simple form: Let r_0 denote the rank deficiency of the covariance matrix of the joint process $(x_t^{(1)\prime}\ x_t^{(2)\prime})'$ and let δ denote a random variate which is 1/2 times the difference of two independent chi-square variates with one degree of freedom. Then S has the distribution of a sum of $r^{(1)}-r_0$ independent variates which have the distribution of δ. If $r_0=r^{(1)}$, then $S=0$.) Findley and Wei (1988) give examples to show that when regressors are equivalent but incomplete, the limiting distribution need not be symmetric.

In the situations (AP1) – (AP3) below, the regressor process $x_t^{(1)}$ would usually be preferred _a priori_ over $x_t^{(2)}$. Note that in (AP1) and (AP2), the regressors are _not_ constrained to be _nested_, that is, the component variables of one need not be linear functions of the other. Also, in (AP1) and (AP3), the regressors are _not_ required to be _complete_; that is, the regression equations need only represent approximations.

(AP1). _The asymptotic error covariance matrices are such that_ $|\Sigma^{(1)}|<|\Sigma^{(2)}|$.

Under (AP1), both (MAIC) and (EMP) prefer $x_t^{(1)}$ with probability approaching 1 as $N \rightarrow \infty$. This will follow from

$$(2.3) \qquad DAIC_N^{(1,2)} \rightarrow -\infty \text{ and } \hat{E}_N^{(1)} - \hat{E}_N^{(2)} \rightarrow \infty$$

with probability 1.

Since (AP1) holds, in particular, if $x_t^{(1)}$ is complete and $x_t^{(2)}$ is not, this result generalizes the well-known fact that, asymptotically, (MAIC) will prefer a fully parameterized regression to an underparameterized one, with probability 1.

(AP2). Both regressor processes $x_t^{(1)}$ and $x_t^{(2)}$ are complete, but $r^{(1)} = \dim x_t^{(1)} < r^{(2)} = \dim x_t^{(2)}$.

Under (AP2), it can be shown that

$$(2.4) \qquad \hat{E}_N^{(1)} - \hat{E}_N^{(2)},\ \hat{L}_N^{(2)} - \hat{L}_N^{(1)} \rightarrow_{\text{dist.}}$$

$$S + (1/2)\ \chi^2_{q(r^{(2)} - r^{(1)})}$$

where the right-hand side denotes the distribution of a variate which is the sum of two independent components, one having the symmetric distribution of the variate S described after (2.2) and the other the distribution of 1/2 times a chi-square variate with $q(r^{(2)} - r^{(1)})$ degrees of freedom. It follows from (2.4) that

(2.5) $$\mathrm{DAIC}_N^{(1,2)} \xrightarrow[\text{dist.}]{} 2S + \{\chi^2_{q(r^{(2)}-r^{(1)})} - 2q(r^{(2)}-r^{(1)})\}$$

It can be shown that var(S)=0 if and only if $x_t^{(1)}$ is a linear transform of $x_t^{(2)}$, in which case the comparison is a nested one. Thus, (2.5) shows that $\mathrm{DAIC}_N^{(1,2)}$ will have greater variance with non–nested comparisons than with nested comparisons. Since $\mathrm{DAIC}_N^{(1,2)}$ behaves completely differently asymptotically under (AP1) than under (AP2), the sampling properties implied by (2.5) can only be assumed if one is certain that (AP2) holds. But in this case, one would want to use a regressor process which was a linear transform of both $x_t^{(1)}$ and $x_t^{(2)}$ instead of using these processes. For these reasons, (2.5) is not useful for tests of hypotheses.

A collateral result which is more useful concerns the convergence of the expected values under (AP2):

(2.6) *The means* $E\{\hat{L}_N^{(2)} - \hat{L}_N^{(1)}\}$, $(-1/2)E\{\mathrm{DAIC}_N^{(1,2)}\}$, and $E\{\hat{E}_N^{(1)} - \hat{E}_N^{(2)}\}$ *all converge to* $(1/2)q(r^{(2)} - r^{(1)})$ as $N \rightarrow \infty$.

Since $r^{(2)} - r^{(1)}$ is positive, this shows that (EMP) *and* (MAIC) *prefer the more parsimonious model with regressor* $x_t^{(1)}$, *on average, as* $N \rightarrow \infty$. It also follows from (2.6) that $\mathrm{DAIC}_N^{(1,2)}$ is an *asymptotically unbiased* estimator of minus twice the entropy difference. This is Akaike's famous result, for which a complete proof has heretofore been lacking in the stochastic regression context. Findley and Wei (1988) present a formula which shows that the *mean* of $\hat{E}_N^{(1)} - \hat{E}_N^{(2)} - \mathrm{DAIC}_N^{(1,2)}$ tends continuously to 0 as the situation changes from (AP1) to (AP2); that is, the mean will be close to zero if the regressors are almost correct. This stability may help to explain (MAIC)'s effectiveness.

The convergence–in–mean result, (2.6), is easy to motivate from (2.4), but quite difficult to prove rigorously. The first complete verification is given in Findley and Wei (1988) for the case of (subvectors of) Gaussian autoregressive processes. The convergence–in–probability or –distribution results discussed above can be shown to hold under quite general assumptions which encompass the most familiar classes of stationary and nonstationary time series processes.

Finally, we consider the nested situation, with not necessarily complete regressors. Let us define a regressor process z_t to be *superfluous given* x_t if the processes x_t and $(x_t', z_t')'$ are equivalent.

(AP3). *The regressor process* $x_t^{(2)}$ *has the form* $x_t^{(2)} = (x_t^{(1)\prime}\ \ z_t')'$, *where* z_t *is superfluous given* $x_t^{(1)}$.

Under (AP3), we will show that

$$\hat{E}_N^{(1)} - \hat{E}_N^{(2)},\ \hat{L}_N^{(2)} - \hat{L}_N^{(1)} \to_{\text{dist.}} \sum_{i=1}^{q(r^{(2)}-r^{(1)})} \lambda_i^2 \chi_{1,i}^2 \tag{2.7}$$

where the distribution on the right–hand side is that of a sum of $q(r^{(2)} - r^{(1)})$ independent variates which are positive multiples of a chi–square variate with 1 degree of freedom. Thus, in this situation, (EMP) *prefers* $x_t^{(1)}$ *with probability approaching* 1 *as* $N \to \infty$. However, $(-1/2)\mathrm{DAIC}_N^{(1,2)}$ now has an asymptotic distribution whose mean is $q(r^{(2)} - r^{(1)}) - \sum_{i=1}^{q(r^{(2)}-r^{(1)})} \lambda_i^2$, and it is shown in Findley and Wei (1988) that this mean will be positive if $x_t^{(1)}$ is close to complete.

The situation (AP3) is, however, one in which the a priori assumption, that the more parsimonious model is better, is not valid in complete generality. By considering regressors which cannot reproduce the data mean, violating an assumption made concerning (1.1), a counterexample is obtained in section 2c. below.

2b. More on the distribution of (2.5). The asymptotic distribution of $DAIC_N^{(1,2)}$ described in (2.5) is not a practically useful one, since it assumes that both $x_t^{(1)}$ and $x_t^{(2)}$ have linear transformations which yield correct regressions. It does, however, shed some light on how the presence of entries in both regressions which are not linear transforms of the other regressor diminishes the probability of DAIC's selecting the more parsimonious model in comparison to the situation in which $x_t^{(1)}$ is nested in $x_t^{(2)}$. If we define r_0 as before (below (2.2)) and set $m = q(r^{(1)}-r_0)$ and $d = q(r^{(2)}-r^{(1)})$, then with x_{m+d}^2 and x_m^2 denoting independent chi-square variates of degrees m+d and m, respectively, we obtain for (2.5) that

$$\lim_{N\to\infty} P(DAIC_N(1,2) < 0) = P(x_{m+d}^2 - x_m^2 < 2d),$$

some values for which are tabled below.

Table 2.1

$$P_{m,d} = P(\chi^2_{m+d} - \chi^2_m < 2d)$$

(From a sample of 50,000 pairs of chi–square pseudorandom deviates)

			d				
m	1	2	6	12	18	...	∞
0	.84	.86	.94	.98	.99		1.0
1	.74	.81	.92	.98	.99		1.0
2	.68	.76	.91	.97	.99		1.0
6	.59	.67	.85	.95	.98		1.0
12	.56	.62	.79	.92	.97		1.0
18	.55	.60	.75	.89	.96		1.0
∞	.50	.50	.50	.50	.50		

Conjectures:

$$P_{m+1,d} < P_{m,d}$$

$$P_{m,d} < P_{m,d+1}$$

2c. Related Criteria and Two Counterexamples to Parsimony: The results (2.2) – (2.6) can be shown to similarly support the use of Schwarz's minimum BIC criterion, the Hannan and Quinn HQ criterion and the cross–validation criteria of Stoica et al. (1986). (The BIC and HQ statistics have the form of $D_N^{(1,2)}$ below, with $C_N = (1/2) \log N$, respectively $(2 + \epsilon) \log \log N$, for any $\epsilon>0$.) The Schwarz and Hannan–Quinn criteria have an additional property not possessed by MAIC of preferring the model with fewer parameters, with probability approaching one as $N \rightarrow \infty$, when two equivalent but not complete regressors are being compared.

Indeed, the fact that, in this situation, $\hat{L}_N^{(1)} - \hat{L}_N^{(2)}$ is bounded in probability means that any criterion of the form

$$D_N^{(1,2)} = -2\{\hat{L}_N^{(1)} - \hat{L}_N^{(2)}\} + 2C_N\{q(r^{(1)} - r^{(2)})\}$$

with $C_N \rightarrow \infty$ will have the property that

$$P(D_N^{(1,2)} < 0) \rightarrow 1 ,$$

leading to a consistent preference for the more parsimonious regressor. This property is not always desirable, as the following two examples show.

Counterexample to No. 1. To begin generally, suppose that a covariance stationary time series with mean 0, whose autocovariance sequence $\Gamma_k = Ey_t y_{t-k}$ and autocorrelation sequence $\rho_k = \Gamma_k/\Gamma_0$ are absolutely summable. Consider the regressors $x_t^{(1)} = y_{t-2}$ and $x_t^{(2)} = [y_{t-1} \ y_{t-3}]'$. The associated idealized regressions satisfying (1.2) are

$$y_t = A_2^{(1)} y_{t-2} + e_t^{(1)} , \qquad \text{Model 1}$$

with $A^{(1)} = \rho_2$, and

$$y_t = A_1^{(2)} y_{t-1} + A_3^{(2)} y_{t-3} + e_t^{(2)} \qquad \text{Model 2}$$

with

$$A_1^{(2)} = \frac{\rho_1 - \rho_2\rho_3}{1-\rho_2^2} \tag{2.8}$$

and

$$A_3^{(2)} = \rho_3 - \rho_2 A_1^{(2)} . \tag{2.9}$$

If, as we assume hereafter,

$$\rho_1 = \rho_2 = \rho_3 = 0 \tag{2.10}$$

then $A_2^{(1)} = A_1^{(2)} = A_3^{(2)} = 0$, so that the regressors $x_t^{(1)}$ and $x_t^{(2)}$ are equivalent given y_t.

Consider the situation wherein the regressions estimated for $y_1,\ldots,y_N$ are used to estimate (predict) y_t from an independent replicate (realization) $\tilde{y}_t$ of the y_t process. The prediction errors are

$$\bar{e}_t^{(1)} = \tilde{y}_t - \hat{A}_{N,2}^{(1)} \tilde{y}_{t-2} \tag{2.11}$$

and

$$\bar{e}_{N+1}^{(2)} = \tilde{y}_t - \hat{A}_{N,1}^{(2)} \tilde{y}_{t-1} - \hat{A}_{N,3}^{(2)} \tilde{y}_{t-3} \ . \tag{2.12}$$

Let $\tilde{E}$ denote the expectation operator for the $\tilde{y}_t$ process. Since the terms on the right hand side of (2.11) and (2.12) are mutually $\tilde{E}$–uncorrelated, because of (2.11) and the independence of the y_t and $\tilde{y}_t$ processes, the mean square prediction errors are given by

$$E\tilde{E}(\bar{e}_t^{(1)})^2 = \Gamma_0 + E(\hat{A}_{N,2}^{(1)})^2\Gamma_0$$

and

$$E\tilde{E}(\bar{e}_t^{(2)})^2 = \Gamma_0 + E(\hat{A}_{N,1}^{(2)})^2\Gamma_0 + E(\hat{A}_{N,3}^{(2)})^2\Gamma_0 \ .$$

We will demonstrate a Gaussian process y_t satisfying (2.10) and also

$$E(\hat{A}_{N,1}^{(2)})^2 + E(\hat{A}_{N,3}^{(2)})^2 < E(\hat{A}_{N,2}^{(1)})^2 \tag{2.13}$$

for N sufficiently large, from which it follows that model 2, with two superfluous parameters, has smaller mean square prediction error than the more parsimonious model 1. The result (7.16) of Findley and Wei (1986) combined with (VI.3.12) of Hannan (1970) shows that $NE(A_{N,k}^{(i)})^2$ converges to the asymptotic variance of the limiting distribution of $N^{1/2}A_{N,k}^{(i)}$. This variance can be obtained from the central

Limit Theorem of (6) or by observing that the asymptotic distribution is that of the sample autocorrelation at lag k (cf. (2.8) and (2.9)), so that Bartlett's formula, given as (47) on p. 488 of Anderson (1971), applies:

$$\lim_{N\to\infty} NE_N(\hat{A}_{N,k}^{(i)})^2 = \sum_{j=-\infty}^{\infty} (\rho_j^2 + \rho_{j+k}\rho_{j-k}) \tag{2.14}$$

To generate a variety of autocorrelation sequences, it is simplest to specify partial autocorrelations ϕ_{kk} and obtain the autocorrelations from the Yule–Walker equations and the Levinson–Durbin algorithm. The condition (2.10) is equivalent to $\phi_{11} = \phi_{22} = \phi_{33} = 0$, and the sign changes needed in the $\rho_{j+k}\rho_{j-k}$ terms of (2.14) to obtain (2.13) are obtained by mixing the signs and magnitudes of ϕ_{kk}, $k>3$ appropriately. Armed with this strategy, my associate Marian Pugh found, for example, that $\phi_{44} = \phi_{55} = .3$, $\phi_{66} = -.6$, $\phi_{kk} = 0$, $k\geq 7$, yields $NE(\hat{A}_{N,2}^{(1)}) \doteq 4.38$, $NE(\hat{A}_{N,1}^{(2)}) \doteq 2.04$ and $NE(\hat{A}_{N,3}^{(2)}) \doteq .6$ from which (2.13) follows.

In contrast to the behavior of BIC and other statistics of the form $D_N^{(1,2)}$ above, Akaike's Entropy Maximization Principle prefers the regressor $x_t^{(2)}$, on average, as $N\to\infty$, in the following sense: if we denote the asymptotic variance (2.14) by V(k), then it follows from formula (3.4) of Findley (1985) that, under (2.10),

$$\lim_{N\to\infty} E(\hat{E}_N^{(1)} - \hat{E}_N^{(2)}) = (1/2)\ (V(1) + V(3) - V(2)) \ ,$$

which is <u>negative</u> for our example. Also, using (2.1) and the limiting distribution given below for $DAIC_N^{(1,2)}$, one can show that $P(\hat{E}_N^{(2)}-\hat{E}_N^{(1)} > 0)$ converges to 0.56.

Since

$$\lim_{N\to\infty} E(DAIC_N^{(1,2)}) = V(1) + V(3) - V(2) - 2$$

is negative, Akaike's minimum AIC criterion prefers the less successful regressor $x_t^{(1)}$, on average, as N→∞. The limiting distribution of DAIC is (–2) + a linear combination of three independent χ_1^2 variates,

$$(-2) + (-4.37)\chi_{1,1}^2 + (2.48)\chi_{1,2}^2 + (0.15)\chi_{1,3}^2 .$$

(The coefficients are the eigenvalues of the matrix C_A^T diag(1,–1,1)C_A, where $C_A C_A^T$ is the Cholesky factorization of the asymptotic covariance matrix of $\hat{A}_{N,1}^{(2)}$, $\hat{A}_{N,2}^{(1)}$, $\hat{A}_{N,3}^{(2)}$.) As a consequence, $DAIC_N^{(1,2)}$ is positive, and prefers $x_t^{(2)}$, with large–sample probability 0.2. To this limited extent, MAIC outperforms $D_N^{(1,2)}$–based criteria.

Counterexample No. 2. We now present a second, more elementary, example, describing an incorrect, superfluous, fixed regressor which offers improved predictive performance over the model which omits it.

Suppose that for some C>1,

$$y_t = 2C\sigma_e t^{-1/2} + e_t \text{ , } t = 1,2,\ldots$$

where e_t ~ IID$(0,\sigma_e^2)$, and consider the fitting of a constant mean to N observations $y_1,\ldots,y_N$, using $x_t^{(1)}\equiv 1$, so that $\hat{A}_N = N^{-1}\Sigma_{t=1}^N y_t$. Since $\hat{A}_N \to 0$ (w.p.1), this regressor is superfluous. However, if $\tilde{y}_1,\ldots,\tilde{y}_N$ denote an independent replicate, it is easy to check that

$$\lim_{N\to\infty}\{E\tilde{E}\sum_{t=1}^{N}(\tilde{y}_t-\hat{A}_N)^2 - \tilde{E}\sum_{t=1}^{N}\tilde{y}_t^2\} = \sigma_e^2(1 - C^2) ,$$

a negative quantity. Hence the regressor process $x_t^{(1)}$ has predictive value over the null regressor $x_t^{(0)} \equiv 0$, when N is sufficiently large.

We now determine the preferences of the various model selection criteria. Set $\hat{\sigma}_N^2 = N^{-1}\Sigma_{t=1}^N(y_t-\hat{A}_N)^2$, $\hat{\sigma}_{0,N}^2 = N^{-1}\Sigma_{t=1}^N y_t^2$, $\hat{L}_N^{(0)} = L_N[\hat{\sigma}_{0,N},0]$, $\hat{L}_N^{(1)} = L_N[\hat{\sigma}_N,\hat{A}_N]$ and define $\hat{E}_N^{(0)}$ and $\hat{E}_N^{(1)}$ analogously. Let $\chi_1^2(C^2)$ denote a chi-square variate with 1 d.f. and non-centrality parameter C^2. Observing that $N^{1/2}(\hat{A}_N/\sigma_e)\to_{\text{dist.}}N(C,1)$, it is straightforward to verify that

$$-2\{\hat{L}_N^{(0)} - \hat{L}_N^{(1)}\}\to_{\text{dist.}} \chi_1^2(C^2)$$

Thus , from our previous discussion, the statistics $D_N^{(0,1)}$ will have a consistent preference for the more parsimonious model (with $x_t^{(0)}$). On the other hand,

$$\text{DAIC}_N^{(0,1)}\to_{\text{dist.}} \chi_1^2(C^2) - 2 ,$$

which has positive mean C^2-1, as does the limiting distribution of twice the entropy difference,

$$2\{\hat{E}_N^{(1)} - \hat{E}_N^{(0)}\}\to_{\text{dist.}} C^2 - \chi_1^2 .$$

Thus, in an average sense, MAIC and EMP prefer the regressor having more predictive power, $x_t^{(1)}$. Also, by choosing C sufficiently large, the probability of this choice can be made arbitrarily close to 1.

Since the distributions $\chi_1^2 - C^2$ and $\chi_1^2(C^2)$ are different, this example reveals, too, that results like (2.1) and (2.7) depend on the assumption implicit in the conditions imposed on (1.1) that the mean of y_t can be obtained as a linear function of the components of x_t.

2d. Additional Discussion. The result (2.1) shows that $(-1/2)\mathrm{DAIC}_N^{(1,2)}$ is not a consistent estimator of $\hat{E}_N^{(1)} - \hat{E}_N^{(2)}$ under (AP2). Indeed, the fact that, asymptotically, $\hat{L}_N^{(1)} - \hat{L}_N^{(2)}$ has the same sign as $\hat{E}_N^{(1)} - \hat{E}_N^{(2)}$ for regressors associated with different entropies, but the opposite sign for equivalent regressors, suggests that an estimator of $\hat{E}_N^{(1)} - \hat{E}_N^{(2)}$ cannot be found which will be consistent in both situations.

Regressors being compared in applications are likely to be neither complete nor equivalent, but might come close enough to having these properties that, for the given sample size N, $E_{N,\infty}^{(1)} - E_{N,\infty}^{(2)}$ does not dominate the behaviors of $\hat{E}_N^{(1)} - \hat{E}_N^{(2)}$ and $\hat{L}_N^{(2)} - \hat{L}_N^{(1)}$, and, also, that the means of $\{\hat{L}_N^{(1)} - L_{N,\infty}^{(1)}\} - \{\hat{L}_N^{(2)} - L_{N,\infty}^{(2)}\}$ and $\{E_{N,\infty}^{(1)} - \hat{E}_N^{(1)}\} - \{E_{N,\infty}^{(2)} - E_N^{(2)}\}$ are close to $(1/2)q(r^{(1)} - r^{(2)})$. For this situation, an alternative approach to DAIC suggests itself which reveals a possible direction for further research. Consider the problem of finding a constant C such that

$$Q_N = \frac{\hat{L}_N^{(1)} - \hat{L}_N^{(2)} - C}{\hat{E}_N^{(1)} - \hat{E}_N^{(2)}}$$

is positive as often as possible. (Then we can use the sign of the numerator to estimate the sign of the denominator.) We can rewrite Q_N in the form

$$Q_N = 1 - \delta_N/(\hat{E}_N^{(1)} - \hat{E}_N^{(2)}),$$

with

$$\delta_N = \{\hat{E}_N^{(1)} - \hat{E}_N^{(2)}\} - \{\hat{L}_N^{(1)} - \hat{L}_N^{(2)} - C\} .$$

From the discussion after (LD) and (ED), the dominant term in δ_N is

$$\begin{aligned} &\{L_{N,\infty}^{(1)} - E_{N,\infty}^{(1)}\} - \{L_{N,\infty}^{(2)} - E_{N,\infty}^{(2)}\} \\ &= (-1/2)\{\mathrm{tr}\Sigma^{(1)-1}(\sum_{t=1}^{N} e_t^{(1)} e_t^{(1)\prime}) \\ &\quad - \mathrm{tr}\Sigma^{(2)-1}(\sum_{t=1}^{N} e_t^{(2)} e_t^{(2)\prime})\}, \end{aligned}$$

which has order $N^{1/2}$ and mean zero. One strategy might be to choose $C(=C_N)$ in such a way that $EQ_N \doteq 1$, but determining the required expression for such a C seems very difficult. A simpler, related strategy is to choose C so that $E\delta_N \doteq 0$, which is accomplished for our situation by Akaike's choice $C = q(r^{(1)} - r^{(2)})$. (since the sign of $\hat{E}_N^{(1)} - \hat{E}_N^{(2)}$ is unknown, we do not have a preferred sign for δ_N.) Thus, from this perspective, Akaike's criterion seems more intelligible than those of Schwarz or Hannan and Quinn. Deeper analyses might support other strategies. Their appeal would depend on the simplicity and stability of the resulting formulas for C.

2e. Comparing Both Regressors and Error Covariance Structures: As we shall illustrate in section 3, there are situations in which the models being compared differ not only in the choice of regressors but also according to whether or not a simplifying block–diagonal structure is assumed for the error covariance matrix,

$$\Sigma^{(x)} = \begin{bmatrix} \Sigma & 0 \\ 0 & \bar{\Sigma} \end{bmatrix} , \tag{2.8}$$

which reduces the number of error covariances estimated. The discussion of (AP1) and (2.3) is not affected by such model differences. Also, if the form of (2.8) is correct (but not necessarily assumed), so that estimates of error covariances in the off–diagonal blocks are superfluous, and if the regressand and regressor processes are jointly Gaussian, then the rest of the discussion about a priori conditions and their consequences generalizes in an obvious way: a model with smaller total number of estimated parameters (coefficients, error covariances) is preferred, and the magnitude of the difference of the number of parameters estimated replaces the quantity $q(r^{(2)} - r^{(1)})$. For non–Gaussian error processes, the same results hold provided the fourth–cumulant terms associated with the assumed–zero entries of $\Sigma^{(x)}$ are zero, see (3.4) below.

We will now illustrate the kinds of comparisons discussed above with some analyses related to modeling for the design of a statistical autopilot for a ship.

3. Ship Autopilot Modeling: Amerika Maru Data

3a. Regressor Selection: In Kitagawa and Ohtsu (1976) and Ohtsu et al. (1979) and the papers referenced there, the design and testing of a stochastic–regression–model–based ship autopilot is described. The success of this experiment influenced the design of a new ship (Shoji Maru III) incorporating such an autopilot (personal communication from K. Ohtsu). The principle variable to be controlled is yaw (Y), the angular deviation of the ship's forward movement from the intended direction, measured at the bridge. Other less important but useful variables to control include roll (R) and pitch (P). The rudder angle (RU) is the main controller input variable, but measured values of the lateral acceleration (YACC) and vertical acceleration (ZACC) of the forepeak may also provide useful information for the controller/autopilot.

Our first analysis will seek to determine the situations in which ZACC is a useful controller input variable for a specific ship: we consider the problem of choosing between the regressors $x_t^{(m)}$ and $\tilde{x}_t^{(M)}$, these being defined by

$$x_t^{(m)} =$$

$$(Y_{t-1}\ R_{t-1}\ P_{t-1}\ RU_{t-1}\ YACC_{t-1}$$

$$\cdots Y_{t-m}\ R_{t-m}\ P_{t-m}\ RU_{t-m} YACC_{t-m})'$$

and

$$\tilde{x}_t^{(M)} = (x_t^{(M)\prime}\ ZACC_{t-1} \cdots\ ZACC_{t-M})'$$

for $1 \leq m, M \leq 10$. The modeling will be done with N=894 observations made at 1 second intervals on the container ship Amerika Maru under manual control. These data are discussed in the papers cited above. The bias–adjusted maximized log likelihood value

$$L_N^{adj} = L_N[\hat{\Sigma}_N, \hat{A}_N] - \dim A$$

is used to define AIC = $(-2)L_N^{adj}$. If $\hat{m}$ and $\hat{M}$ denote the lags associated with minimum AIC values for the regressors $x_t^{(m)}, 1 \leq m \leq 10$ and $\tilde{x}_t^{(M)}$, $1 \leq M \leq 10$, respectively, and if $\bar{L}_N^{adj}$ and $\tilde{L}_N^{adj}$ denote the corresponding values of L_N^{adj}, then the use of ZACC in the control model seems worth considering seriously only when

$$DAIC_N = (-2)\{\tilde{L}_N^{adj} - \bar{L}_N^{adj}\}$$

is negative. Results obtained from the program MULCON of Akaike et al. (1985) for seven choices of the regressand y_t are included in Table 3.1 below. The choices for y_t are: Y_t, R_t, P_t, $(Y_t \; R_t)'$, $(Y_t \; P_t)'$, $(R_t \; P_t)'$ and $(Y_t \; R_t \; P_t)'$. In the table, LAG denotes $\hat{M}$ or $\hat{m}$, as appropriate, and

$$\Delta \dim A = q(6\hat{M} - 5\hat{m}) \; ,$$

with q = dim y. The results are consistent: the use of ZACC is favored only when P is one of the controlled variables. This conclusion has engineering plausibility: ZACC is closely related to P but not to the other controlled variables. Thus, MAIC has functioned quite satisfactorily. (Note also that in two cases, $y_t = Y_t$ and $y_t = P_t$, the comparison is between non-nested regressors.)

3b. Coupled Models versus Uncoupled Models: Now we consider another kind of model comparison which sometimes plays a role in regressor selection in control problems. It is a useful if a noncontrolled variable such as ZACC has the property that its innovations (or noise) process is uncorrelated with the innovations processes of the controlled variables. This facilitates the interpretation of the cross–spectrum diagnostics of Otomo et al. (1972) and seems empirically to be associated with some robustness of a model–based controller against the changes which occur in the joint stochastic properties of the measured variables when the controller is implemented.

Let $\bar{y}_t$ consist of the $\bar{q}$ noncontrolled variables whose innovations are under investigation and let $\bar{x}_t$ denote the $\bar{r}$–dimensional regressor selected for modeling $\bar{y}_t$. We thus have a pair of models,

$$\begin{aligned} y_t &= Ax_t + e_t \\ \bar{y}_t &= \bar{A}\,\bar{x}_t + \bar{e}_t \end{aligned} \tag{3.1}$$

Table 3.1. DAIC Values.

	Without ZACC			With ZACC				
y	LAG	dimA	AIC_{894}	LAG	dimA	AIC_{894}	ΔdimA	$DAIC_{894}$
Y,R,P	8	108	18683.	6	90	18693.	18	–10.
Y,R	6	72	12244.	6	60	12236.	12	8.
Y,P	7	84	12468.	7	70	12483.	14	–15.
R,P	8	96	13033.	8	80	13036.	16	–3.
Y	4	24	6027.	8	30	6025.	–6	2.
R	5	30	6574.	5	25	6569.	5	5.
P	7	42	6477.	8	40	6491.	2	–14.
ZACC	3	18	7858.	–	–	–	–	–
Y,R,P,ZACC	6	144	26552.	–	–	–	–	–

and we wish to know if they are uncoupled, meaning that

$$\Sigma_{e\bar{e}} = Ee_t\bar{e}_t{}' = 0 \tag{3.2}$$

holds. By defining entries of the associated coefficient matrix appropriately, (3.1) could be rewritten in a single equation as a regression of $(y_t' \ \bar{y}_t')'$ on a combined regressor, $x_t \vee \bar{x}_t$, which encompasses the contributions of x_t and $\bar{x}_t$. Under (3.2), the covariance matrix of $(e_t \ \bar{e}_t)'$ is block-diagonal,

$$\mathrm{cov}(e_t,\bar{e}_t) = \begin{bmatrix} \Sigma & 0 \\ 0 & \bar{\Sigma} \end{bmatrix} ,$$

where $\bar{\Sigma}$ denotes a model covariance matrix for $\bar{e}_t$. As a consequence, the log likelihood function for the data $(y_t \ \bar{y}_t)'$, $x_t \vee \bar{x}_t$, $t=1,\ldots,N$ has the decomposition

$$L_N[\Sigma,\bar{\Sigma},A,\bar{A}] = L_N[\Sigma,A]+L_N[\bar{\Sigma},\bar{A}] \ . \tag{3.3}$$

Since the results of the previous section apply to the log-likelihood functions on the right hand side of (3.3), analogous results hold for $L_N[\Sigma,\bar{\Sigma},A,\bar{A}]$ and its expected value. Under a condition to be discussed ((3.4) below), the maximum likelihood model associated with (3.3) can be compared to that obtained from the regression of $(y_t \ \bar{y}_t)'$ on some regressor process $\tilde{x}_t$ (which might differ from $x_t \vee \bar{x}_t$), whose associated error covariance matrix is not constrained, and therefore has $q\bar{q}$ additional covariances to be estimated. Let $\tilde{L}_N^{adj}$, L_N^{adj} and $\bar{L}_N^{adj}$ denote the adjusted maximum likelihood values obtained from the least squares regressions of $(y_t \ \bar{y}_t)'$, y_t and $\bar{y}_t$, respectively, on their selected regressors. Let $\tilde{E}_N$, E_N and $\bar{E}_N$ denote the corresponding entropy values. Consider the estimator

$$\mathrm{DAIC}_N = -2(\tilde{L}_N^{adj} - L_N^{adj} - \bar{L}_N^{adj} - q\bar{q})$$

of $-2(\tilde{E}_N - E_N - \bar{E}_N)$. Under (3.2), *and the assumption that the cross-fourth- cumulant terms vanish,*

$$E(\Sigma_{ee}^{-1/2} e_{jt})^2 (\Sigma_{\bar{e}\bar{e}}^{-1/2} \bar{e}_{kt})^2 - 1 = 0 \tag{3.4}$$

$$(1 \le j \le q,\ 1 \le k \le \bar{q}J \quad ,$$

the appropriate versions of formulas (1.8) and (2.5), which are the results discussed in subsection 2c, show that DAIC_N is a bias-corrected estimator of $-2(\tilde{E}_N - E_N - \bar{E}_N)$. Thus, if DAIC_N is positive, the uncoupled models (3.1) are preferred. We will consider two examples, assuming (3.4) holds for both comparisons in order to facilitate the discussion. (Estimators of the left-hand side of (3.4) are highly variable, so this is a more difficult condition to analyze than (3.2)!).

In the *Amerika Maru* analysis for M=6, the regression residuals of ZACC_t and P_t had the smallest sample correlations with the other residuals. The calculated values are given in Table 3.2.

Table 3.2.

Sample Correlations of Regression Residuals.

Residual	Y	R	P
ZACC	0.08	0.02	−0.01
P	0.21	0.18	1.00

The analysis of the validity of (3.2) for $y_t = (Y_t \ R_t \ P_t)'$ and $\bar{y}_t = ZACC_t$ is the appropriate one for our earlier discussion. We obtain from Table 3.1 that

$$DAIC_{894} = 26552.-18683.-7858. + 6. = 5.0 ,$$

favoring (3.2). To demonstrate a contrasting result, we examine $y_t = (y_t \ R_t)'$ and $\bar{y}_t = P_t$. From Table 3.1 again,

$$DAIC_{894} = 18683.-12244.-6477. + 4. = -34.0 ,$$

rejecting (3.2).

A traditional statistical test of (3.2) would involve (assuming (3.4) and) forcing the comparison to be a nested one by considering only a single regressor, so that $\tilde{x}_t=x_t=\bar{x}_t$ (violated in both of our examples and in the example of Kitagawa and Ohtsu (1976)). Then the values of

$$N\{\log|\hat{\Sigma}_{ee}||\hat{\Sigma}_{\bar{e}\bar{e}}| - \log|\hat{\Sigma}|\} ,$$

with $\hat{\Sigma}$ defined by

$$\hat{\Sigma} = \begin{vmatrix} \hat{\Sigma}_{ee} & \hat{\Sigma}_{e\bar{e}} \\ \hat{\Sigma}'_{e\bar{e}} & \hat{\Sigma}_{\bar{e}\bar{e}} \end{vmatrix} ,$$

would be compared to a preselected critical value of a $\chi^2_{q\bar{q}}$ distribution. The use of AIC offers more flexibility.

Other control applications of MFPEC/MAIC are described in Nakamura et al. (1986) and Otomo et al. (1972). Many more are not publicly documented for company confidentiality reasons. Mr. Kazutsuro Toki of System Soogoo Kaihatsu kindly informed the author in response to a query that his company has been involved in more than 60 commercial implementations based on the statistical methodology described in these papers. The University of Tulsa has distributed more than 700 copies of the TIMSAC software package.

Acknowledgement

Some of the investigations reported here were carried out while the author was a Visiting Professor at the Institute of Statistical Mathematics in Tokyo. This research benefitted substantially from stimulating conversations there with G. Kitagawa and with H. Akaike, R. Shimizu and T. Ozaki. The author is also grateful to E. Arahata and Marian Pugh for computing assistance. Finally, I wish to express my gratitude to W. W. Davis for useful comments on an earlier version of this paper.

References

Akaike, H. (1971). "Autoregressive Model Fitting for Control," Ann. Inst. Statist. Math. 23, 163–180.

Akaike, H. (1973). "Information Theory and an Extension of the Likelihood Principle," in 2nd International Symposium on Information Theory, eds. B. N. Petrov and F. Czaki, Budapest: Akademia Kiado, 267–287.

Akaike, H., T. Ozaki, M. Ishiguro, Y. Ogata, G. Kitagawa, Y–H Tamura, E. Arahata, K. Katsura, and Y. Takura (1985), TIMSAC 84, Part 2, Computer Science Monographs No. 23, Tokyo: Inst. Statist. Math.

Anderson, T. W. (1971). "The Statistical Analysis of Time Series," New York: Wiley.

Chan, N. H. and C. Z. Wei (1988), "Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes," Ann. Stat. 16, 367–401.

Findley, D. F. (1985). "On the Unbiasedness Property of AIC for Exact or Approximating Linear Stochastic Time Series Models," J. Time Ser. Anal. 6, 229–252.

Findley, D. F. and C. Z. Wei (1986). "An Analysis of AIC for Linear Stochastic Regression and Control," Institute of Statistical Mathematics Research Memorandum No. 32.

Findley, D. F. and C. Z. Wei (1988). "Results on Kullback–Leibler Entropy, AIC and the Comparison of Not Necessarily Nested Time Series Regression Models," in preparation.

Hannan, E. J. (1970). Multiple Time Series, New York: Wiley.

Hosoya, Y. and M. Taniguchi (1982). "A Central Limit Theorem for Stationary Processes and the Parameter Estimation of Linear Processes," Ann. Statist. 10, 132–153.

Kitagawa, G. and K. Ohtsu (1976). "The Statistical Control of Ship's Course Keeping Motion," Proc. Inst. Statist. Math. 23, 105–128. (in Japanese)

Lai, T. L. and C. Z. Wei (1982). "Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems," Ann. Statist., 10, 154–166.

Ljung, L. and P. Caines (1979). "Asymptotic Normality of Prediction Error Estimates for Approximate System Models," Stochastics 3, 29–45.

Nakamura, H., M. Uchida, Y. Toyota and M. Kushihashi (1986). "Optimal Control of Thermal Power Plants," ASME Winter Annual Meetings Proceedings., 86–4A/DSC–14.

Ohtsu, K, M. Horigome and G. Kitagawa (1979). "A New Ship's Autopilot Design Through a Stochastic Model," Automatica, 15, 255–268.

Otomo, T., T. Nakagawa and H. Akaike (1972). "Statistical Approach to Computer Control of Cement Rotary Kilns," Automatica 8, 35–48.

Shimizu, R. (1978). "Entropy Maxamization Principle and Selection of the Order of an Autoregressive Gaussian Process," Ann. Inst. Statist. Math. 30, 263–270.

Stoica, P. and P. Eykhoff, P. Janssen, T. Söderström (1986). "Model Structure by Cross–Validation," Int. J. Control 43, 1841–1878.

Tiao, G. C. and R. S. Tsay (1983). "Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARIMA Models," Ann. Statist. 11, 856–871.