Ann. Inst. Statist. Math.
Vol. 43, No. 3, 505–514 (1991)

COUNTEREXAMPLES TO PARSIMONY AND BIC*

DAVID F. FINDLEY

Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan

(Received February 6, 1990; revised April 1, 1991)

Abstract. Suppose that the log-likelihood-ratio sequence of two models with different numbers of estimated parameters is bounded in probability, without necessarily having a chi-square limiting distribution. Then BIC and all other related "consistent" model selection criteria, meaning those which penalize the number of estimated parameters with a weight which becomes infinite with the sample size, will, with asymptotic probability 1, select the model having fewer parameters. This note presents examples of nested and non-nested regression model pairs for which the likelihood-ratio sequence is bounded in probability and which have the property that the model in each pair with *more* estimated parameters has better predictive properties, for an independent replicate of the observed data, than the model with fewer parameters. Our second example also shows how a one-dimensional regressor can overfit the data used for estimation in comparison to the fit of a two-dimensional regressor.

Key words and phrases: Model selection, linear regression, misspecified models, AIC, BIC, MDL, Hannan-Quinn criterion, overfitting.

1. Introduction

Box and Jenkins ((1976), p. 17) formulated the principle of parsimony in modeling as the use of the "smallest possible number of parameters for adequate representation" of the data. Model adequacy can usually be determined only with the aid of later data which are not available when the model is fitted. So it is likely that most modelers think of this principle as asserting that, given several models which fit the data equally well, the one with fewest estimated parameters should be preferred. This formulation of the principle can be subjected to a mathematical analysis if the notion of fit is made mathematically precise. For linear least squares regression models, a natural measure of fit, which we adopt, is the large-sample limit of the sample variance of the regression residuals. With this measure, two fitted models whose Gaussian log-likelihood-ratio stays bounded in probability as

* An earlier version of this article was presented at the Symposium on the Analysis of Statistical Information held in the Institute of Statistical Mathematics, Tokyo during December 5–8, 1989.

the sample size N increases have the same fit, because the log-likelihood-ratio is proportional to N times the log of the ratio of the sample variances, see below, so this latter ratio must converge to one.

When the log-likelihood-ratio sequence of two models with *different numbers of parameters* is bounded in probability, then model selection criteria like BIC, whose penalty for estimated parameters becomes infinite with N, will obey the principle of parsimony in the *strong* sense that the probability of selecting the model with fewer parameters approaches one as N increases. This property has sometimes been assumed to establish the superiority of such criteria over a criterion like AIC, see Kashyap (1980) and Raftery and Martin (1988), because for AIC, the large-sample probability of selecting a correctly parameterized model instead of an over-parameterized model is a bit less than one, see Shibata (1976) and Woodroofe (1982), for example.

Of course, arguments based upon the assumption of a correct model are somewhat remote from the situation of practicing statistical modelers. This paper demonstrates in a quite elementary fashion that the principle of parsimony formulated above is not generally valid when incorrect models are compared, and it shows that this principle can fail in a way that makes the strong parsimony property of criteria like BIC disadvantageous. An implication of this fact for model selection theory is suggested in Section 5.

Our examples are misspecified regression models for univariate data y_t, $1 \leq t \leq N$, which are estimated, given column vector regressors x_t, $1 \leq t \leq N$, of dimension $dim\, x_t$, by maximizing a Gaussian log-likelihood function,

$$L_N[\sigma^2, A] \equiv -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{N} (y_t - Ax_t)^2.$$

We use $\equiv$ to indicate the definition of a symbol. This maximization leads to the least squares estimates

$$\hat{A}(N) \equiv \sum_{t=1}^{N} y_t x_t' \left(\sum_{t=1}^{N} x_t x_t' \right)^{-1} \tag{1.1}$$

and

$$\hat{\sigma}(N)^2 \equiv N^{-1} \sum_{t=1}^{N} (y_t - \hat{A}(N)x_t)^2,$$

and to the maximized value

$$\hat{L}_N \equiv L_N[\hat{\sigma}(N)^2, \hat{A}(N)] = -\frac{N}{2} \log 2\pi e \hat{\sigma}(N)^2.$$

Several well known regressor selection criteria have the form

$$\mathrm{CRIT}(N) \equiv -2\hat{L}_N + W_N(dim\, x_t + 1), \tag{1.2}$$

where W_N is a sequence of positive numbers. For example, the AIC of Akaike (1973) is obtained when $W_N \equiv 2$, and the BIC of Schwarz (1978) is obtained when

$W_N \equiv \log N$. For the criterion of Hannan and Quinn (1979), $W_N \equiv 2 \log\log N$. When two competing regressor processes $x_t^{(1)}$ and $x_t^{(2)}$ are being compared by (1.2), the one with the smaller criterion value is favored. (Such criteria are admissible, see Takada (1982).) Observe that

$$\begin{aligned}(1.3)\qquad &\mathrm{CRIT}^{(1)}(N) - \mathrm{CRIT}^{(2)}(N)\\ &= N \log\{\hat{\sigma}^{(1)}(N)^2/\hat{\sigma}^{(2)}(N)^2\} + W_N(\mathit{dim}\, x_t^{(1)} - \mathit{dim}\, x_t^{(2)}).\end{aligned}$$

Therefore, if $W_N \rightarrow \infty$ with N, and if the log-likelihood-ratio is bounded in probability,

$$(1.4)\qquad N \log\{\hat{\sigma}^{(1)}(N)^2/\hat{\sigma}^{(2)}(N)^2\} \sim O_p(1),$$

then the regressor with smaller dimension, and therefore fewer estimated coefficients, will be preferred with probability tending to one. Criteria with this property will be called *strongly parsimonious.* The somewhat differently defined MDL and PLS criteria of Rissanen (1978, 1986, 1989) and the FIC criterion of Wei (1991) also have this property.

In the next two sections, examples will be given to demonstrate that such a consistent preference for a more parsimonious model can be undesirable according to a cost function which is a natural measure of prediction error, which we now describe. Let E denote expectation with respect to the joint distribution of the observed data. The measure assumes that the y-values being predicted and the regressors used for prediction are an independent replicate y_t^*, x_t^*, $1 \leq t \leq N$ of the data used to determine $\hat{A}(N)$ (whose expectation operator is denoted by E^*). If x_t is non-stochastic, then $x_t^* = x_t$. This cost function concerns mean square prediction error over both replicates,

$$C^*(N) \equiv E\left\{E^*\left\{\sum_{t=1}^{N}(y_t^* - \hat{A}(N)x_t^*)^2\right\}\right\}.$$

Greater cost means worse predictive performance.

For comparing regressors $x_t^{(1)}$ and $x_t^{(2)}$, we are interested in the ultimate sign of the cost difference,

$$(1.5)\qquad \Delta^*(N) \equiv C^{*(1)}(N) - C^{*(2)}(N).$$

2. A nested comparison with fixed regressors

Suppose that the correct model for y_t is given by

$$(2.1)\qquad y_t = \frac{1}{2}at^{-1/2} + e_t,$$

where $a^2 > 1$ and e_t, $t = 1, \ldots$ is a sequence of independent $\mathcal{N}(0,1)$ variates. A relevant result is

$$(2.2)\qquad \lim_{N\rightarrow\infty} N^{-1/2} \sum_{t=1}^{N} \frac{1}{2} t^{-1/2} = 1,$$

which can be derived by considering the integral of $(1/2)t^{-1/2}$ over $[1, N]$.

For simplicity, we begin by comparing the null regressor $x_t^{(1)} \equiv 0$, which requires no coefficient estimation and is equivalent to the constraint $\hat{A}^{(1)}(N) = 0$, with the constant mean regressor $x_t^{(2)} \equiv 1$. Then $\hat{A}^{(2)}(N) = \bar{y}(N) \equiv N^{-1}(y_1 + \cdots + y_N)$. It follows from (2.1) and (2.2) that, in distribution,

$$N^{1/2}\bar{y}(N) \to \mathcal{N}(a, 1). \tag{2.3}$$

For later reference, we note that

$$\sum_{t=1}^{N} y_t^2 - \sum_{t=1}^{N}(y_t - \bar{y}(N))^2 = N\bar{y}(N)^2. \tag{2.4}$$

For the independent replicate $y_1^*, \dots, y_N^*$ of $y_1, \dots, y_N$, we calculate

$$\begin{aligned}\Delta^*(N) &\equiv E\left\{E^*\left\{\sum_{t=1}^{N} y_t^{*2} - \sum_{t=1}^{N}(y_t^* - \bar{y}(N))^2\right\}\right\} \\ &= 2NE^*\{\bar{y}^*(N)\}E\{\bar{y}(N)\} - NE\{\bar{y}(N)^2\} \\ &= a^2\left(N^{-1/2}\sum_{t=1}^{N}\frac{1}{2}t^{-1/2}\right)^2 - 1,\end{aligned}$$

whose limiting value, $\Delta^*(\infty) = a^2 - 1$, is positive. Thus, when N is large, the use of the regressor $x_t^{(1)}$ results in greater prediction error, as measured by $C^*(N)$, than the use of $x_t^{(2)}$, which, unlike $x_t^{(1)}$, involves a coefficient estimate.

Now we will verify (1.4), with

$$\begin{aligned}\hat{\sigma}^{(1)}(N)^2 &\equiv N^{-1}\sum_{t=1}^{N} y_t^2, \\ \hat{\sigma}^{(2)}(N)^2 &\equiv N^{-1}\sum_{t=1}^{N}(y_t - \bar{y}(N))^2.\end{aligned}$$

In fact, from (2.3) and (2.4) we obtain $N\{\hat{\sigma}^{(1)}(N)^2 - \hat{\sigma}^{(2)}(N)^2\} \sim O_p(1)$. Since, with probability 1,

$$\hat{\sigma}^{(1)}(N)^2, \hat{\sigma}^{(2)}(N)^2 \to 1, \tag{2.5}$$

we can conclude that

$$Z_N \equiv \{\hat{\sigma}^{(1)}(N)^2/\hat{\sigma}^{(2)}(N)^2 - 1\} \sim O_p(N^{-1}). \tag{2.6}$$

Note that with Z_N so defined, we have

$$N \log\{\hat{\sigma}^{(1)}(N)^2/\hat{\sigma}^{(2)}(N)^2\} = N \log(1 + Z_N).$$

From the Taylor expansion

$$\log(1+Z_N) = Z_N - \frac{1}{2}(1+\tilde{Z}_N)^{-1}Z_N^2,$$

where $\tilde{Z}_N$ is between Z_N and 0, and from (2.6), it follows that

$$N\log\{\hat{\sigma}^{(1)}(N)^2/\hat{\sigma}^{(2)}(N)^2\} \sim NZ_N \tag{2.7}$$

and that (1.4) holds. In (2.7) and in (3.5) below, $\sim$ indicates that the difference between the expressions tends to 0 in probability. Since (1.4) holds, strongly parsimonious criteria will select $x_t^{(1)}$ with asymptotic probability 1.

By contrast, the minimum AIC criterion will lead to a choice of the better predicting regressor $x_t^{(2)}$ with a probability which can be made to be as close to 1 as desired by choosing a^2 large enough: indeed, by (2.4), (2.5) and (2.7), the log-likelihood-ratio has the same limiting distribution as $N\bar{y}(N)^2$, which, by (2.3) is non-central chi-square, $\chi_1^2(a^2)$. Hence

$$\lim_{N\to\infty} \text{pr}\{\text{AIC}^{(1)}(N) - \text{AIC}^{(2)}(N) > 0\} = \text{pr}\{\chi_1^2(a^2) > 2\},$$

which ranges between 0.4517 and 1.0 as a^2 ranges between 1.0 and ∞.

There are also other analyses which confirm the superiority of $x_t^{(2)}$. Consider, for example, the *minimum mean total squared error* of the i-th regressor,

$$\text{MSS}^{(i)} \equiv \min_A E\sum^N (y_1 - Ax_1^{(i)})^2 \qquad (i=1,2).$$

It is easily verified that $\lim_{N\to\infty}\{\text{MSS}_N^{(1)} - \text{MSS}_N^{(2)}\} = a^2 > 0$.

With modest additional computational effort, the reader will be able to verify that the same limiting results hold when $x_t^{(1)} \equiv (-1)^t$ and $x_t^{(2)} \equiv [(-1)^t\ 1]'$.

3. Non-nested, incorrect autoregressions

Now we let y_t denote a mean zero, stationary, Gaussian autoregressive process of order 6, with variance $E(y_t^2) = 1$, whose first three partial autocorrelations are zero,

$$\phi_{11} = \phi_{22} = \phi_{33} = 0. \tag{3.1}$$

If we set $\rho_k \equiv E(y_{t+k}y_t)$, $k = 0, \pm 1, \ldots$, then it follows from the Levinson-Durbin algorithm, see Durbin (1960) or Levinson (1946), that (3.1) is equivalent to

$$\rho_1 = \rho_2 = \rho_3 = 0. \tag{3.2}$$

As a consequence, if $z_t \equiv (y_{t-1}\ y_{t-2}\ y_{t-3})'$, then, with probability 1,

$$\lim_{N\to\infty} N^{-1}\sum_{t=1}^N z_t z_t' = I_3 = Ez_t z_t', \tag{3.3}$$

the identity matrix of order 3.

For a given permutation (j_1, j_2, j_3) of $(1, 2, 3)$, we define the regressors

$$x_t^{(1)} \equiv y_{t-j_1}, \qquad x_t^{(2)} \equiv [y_{t-j_2}\ y_{t-j_3}]'.$$

It follows from the formula (1.1) and from (3.3) that the associated coefficient estimates $\hat{A}^{(1)}(N)$ and $\hat{A}^{(2)}(N)$ are consistent estimates of ρ_{j_1} and $[\rho_{j_2}\ \rho_{j_3}]$.

The details omitted from the argument that we now sketch can be found in Findley and Wei (1988, 1991). Set

$$\hat{\rho}_{j_k}(N) \equiv N^{-1} \sum_{t=1}^{N-j_k} y_{t+j_k} y_t$$

and recall from Anderson ((1971), p. 478) that, since $\rho_{j_k} = 0$,

$$\lim_{N\to\infty} NE\{\hat{\rho}_{j_k}^2(N)\} = V_{j_k}, \qquad N^{1/2}\hat{\rho}_{j_k}(N) \xrightarrow[\text{dist.}]{} \mathcal{N}(0, V_{j_k}), \tag{3.4}$$

where, with $f(\lambda)$ denoting the spectral density of y_t, V_{j_k} is given by

$$V_{j_k} \equiv 4\pi \int_{-\pi}^{\pi} \cos^2 \lambda j_k f^2(\lambda)\, d\lambda.$$

Using the formula for $\sum_{t=1}^{N} y_t x_t^{(i)}$, $i = 1, 2$ implied by (1.1),

$$\delta(N) \equiv \sum_{t=1}^{N} (y_t - \hat{A}^{(1)}(N) x_t^{(1)})^2 - \sum_{t=1}^{N} (y_t - \hat{A}^{(2)}(N) x_t^{(2)})^2$$

reduces to

$$\delta(N) = -\hat{A}^{(1)}(N)^2 \sum_{t=1}^{N} x_t^{(1)^2} + \hat{A}^{(2)}(N) \sum_{t=1}^{N} x_t^{(2)} x_t^{(2)\prime} \hat{A}^{(2)}(N)'.$$

By contrast, since y_t and $x_t^{(i)}$, $i = 1, 2$ are uncorrelated, and since (3.3) holds, we obtain that

$$\delta^*(N) \equiv E^* \left\{ \sum_{t=1}^{N} (y_t^* - \hat{A}^{(1)}(N) x_t^{(1)*})^2 - \sum_{t=1}^{N} (y_t^* - \hat{A}^{(2)}(N) x_t^{(2)*})^2 \right\}$$

simplifies to

$$\delta^*(N) = N\hat{A}^{(1)}(N)^2 - N\hat{A}^{(2)}(N)\hat{A}^{(2)}(N)'.$$

Since the quantities $N^{1/2}\hat{\rho}_{j_k}(N)$ are bounded in probability, see (3.4), it follows from (3.3) and these formulas that

$$\delta^*(N) \sim -\delta(N) \sim N\hat{\rho}_{j_1}^2(N) - N\hat{\rho}_{j_2}^2(N) - N\hat{\rho}_{j_3}^2(N). \tag{3.5}$$

The results (3.4) and (3.5) suggest that the limiting value $\Delta^*(\infty)$ of $\Delta^*(N) = E\delta^*(N)$ (see (1.5)) is given by

$$\Delta^*(\infty) = V_{j_1} - V_{j_2} - V_{j_3}, \tag{3.6}$$

whereas $\Delta(\infty) \equiv \lim_{N\to\infty} E\delta(N)$ has the value

$$\Delta(\infty) = -\Delta^*(\infty). \tag{3.7}$$

A complete verification of (3.6) and (3.7) requires a rather subtle argument, see Findley and Wei (1988, 1991).

A Taylor expansion argument analogous to that of Section 2 shows that $N \log(\hat{\sigma}^{(1)}(N)^2/\hat{\sigma}^{(2)}(N)^2) \sim \delta(N)$. Since the quantities $N^{1/2}\hat{\rho}_{j_k}(N)$ converge in distribution, it follows from (3.5) that (1.4) holds. Thus, all strongly parsimonious criteria will have a consistent preference for $x_t^{(1)}$. By (3.6), this will be undesirable for large N whenever

$$V_{j_1} > V_{j_2} + V_{j_3}. \tag{3.8}$$

We examined this inequality for 1000 AR(6) processes determined by choosing ϕ_{44}, ϕ_{55} and ϕ_{66} independently and uniformly in $(-1, 1)$ and imposing (3.1). The integrals defining the V_{j_k}'s were carefully evaluated numerically. The inequality (3.8) was determined to hold for 577 of the 1000 process, for some permutation (j_1, j_2, j_3). For example, with $\phi_{44} = 0.80$, $\phi_{55} = -0.41$, $\phi_{66} = -0.64$ and $(j_1, j_2, j_3) = (2, 1, 3)$, we obtained

$$V_{j_1} = 26.3, \qquad V_{j_2} = 2.9, \qquad V_{j_3} = 2.4,$$

so that (3.8) is satisfied. These different variance values illustrate the principle underlying the existence of such counterexamples to the principle of parsimony: when incorrect models are considered, the costs associated with estimating different coefficients are not always the same, even when the coefficients are negligible asymptotically, see Findley and Wei (1988) for general formulas.

For this example, the asymptotic probability that AIC chooses the better regressor $x_t^{(2)}$ can be shown to be 0.17.

Now we consider (3.7), which describes an *overfitting principle*: worse prediction performance with an independent replicate is precisely matched by an increased (over)fit to the data used for estimation. This is a general principle in that (3.7) can be shown to hold whenever the difference between the optimal mean square fits $\mathrm{MSS}_N^{(1)}$ and $\mathrm{MSS}_N^{(2)}$ (defined at the end of Section 2) vanishes with increasing N. Our examples satisfying (3.8) demonstrate therefore that a one-dimensional regressor can be overfitting in comparison with a two-dimensional regressor. In other words, *parameter parsimony can increase overfitting.*

4. Other situations

The situations considered above, in which the log-likelihood-ratio sequences were bounded in probability, are a small subclass of the situations in which two competing regression models have the same degree of fit asymptotically. Consider the case in which y_t, $x_t^{(1)}$ and $x_t^{(2)}$ are zero-mean, jointly stationary, ergodic time series, with y_t univariate and $x_t^{(1)}$ and $x_t^{(2)}$ multivariate. The least squares coefficient estimates $\hat{A}^{(1)}(N)$ and $\hat{A}^{(2)}(N)$ of the regression of y_t on $x_t^{(1)}$ and $x_t^{(2)}$ will converge to the coefficient vectors $A^{(1)}$ and $A^{(2)}$ given by

$$A^{(i)} = Ey_t x_t^{(i)\prime}(Ex_t^{(i)}x_t^{(i)\prime})^{-1}, \qquad i = 1, 2.$$

The competing regression models will fit the data equally well, according to the measure of fit used earlier, if and only if the error processes

$$e_t^{(i)} \equiv y_t - A^{(i)}x_t^{(i)}, \qquad i = 1, 2$$

have the same variance,

$$E(e_t^{(1)})^2 = E(e_t^{(2)})^2. \tag{4.1}$$

It can be shown, see Findley and Wei (1988), that the log-likelihood-ratio is bounded in probability if and only if the condition

$$e_t^{(1)} = e_t^{(2)} \quad \text{(almost surely, for all } t) \tag{4.2}$$

holds, which is in general stronger than (4.1) for non-nested regressors.

If (4.1) holds, but not (4.2), then the log-likelihood-ratio is of order $N^{1/2}$ in probability. More precisely, it can be shown as in Theorem 8.4 of Findley (1990), see also Theorem 5.1 of Vuong (1989), that $N^{-1/2}\{\hat{L}_N^{(1)} - \hat{L}_N^{(2)}\}$ will have a Gaussian limiting distribution with mean zero. In this situation, adjustments to the log-likelihood-ratio of the form $W_N(dim\, x_t^{(1)} - dim\, x_t^{(2)})$ with $W_N = o(N^{1/2})$ will have negligible effect asymptotically. Hence AIC and BIC will lead to the same choice of regressor in large samples, with each model having asymptotic probability 0.5 of being selected. Therefore finite-sample analyses would be required to distinguish between the properties of these criteria, or analyses like those of Shibata (1980, 1981) in which the model classes become larger with the sample size in such a way that (4.2) is achieved in the limit.

The other situation in which AIC and BIC have identical preferences with large N is where (4.1) fails. (Techniques for detecting this are presented in Findley (1990).) Then the log-likelihood-ratio goes linearly in N towards $+\infty$ or $-\infty$, carrying with it the AIC and BIC differences. Thus (4.2) characterizes the situation where these criteria can exhibit different large-sample behaviors.

5. Concluding remarks

Motivation for the principle of parameter parsimony is closely tied to the situation in which the models being considered are close to the correct model. Although much of the conceptual paradigm of classical mathematical statistics for correct models can be carried over into situations where one avoids the unrealistic assumption that a model class under consideration is correct, see White (1990), our examples reveal that this is not the case with parameter parsimony.

The deep investigations of Shibata (1980, 1981) already showed that strongly parsimonious criteria can lead to regressor selections which lack predictive power relative to AIC's selections in certain situations in which the true model is specified by infinitely many parameters. Our examples address the principle of parsimony more directly and simply than these papers, which are concerned with optimality properties of AIC.

We were motivated to look for these examples by criticisms of AIC like those mentioned in the Introduction and by a train of thought something like "if AIC is so frequently successful in applications and lacks the (strong) parsimony property, then this property must have limited value." It does not follow from our results that a strongly parsimonious criterion cannot be useful. What follows, we suggest, is that if a strongly parsimonious criterion is useful in a variety of applications, its utility will be better explained by some deeper principle which can be formulated without the assumption that one of the models considered is correct. Both Akaike's Entropy Maximization Principle motivating AIC (see Akaike (1985)) and Rissanen's Minimum Description Length Principle (1978, 1986, 1989) have such formulations. Also, Poskitt (1987) has given a derivation of a Bayesian decision-theoretic criterion in this spirit which penalizes for parameter estimation similarly to BIC ($W_N = \log N + O_p(1)$) and which might therefore help to explain good performance by BIC.

Acknowledgements

The computer programming required for the examples described in Section 3 was done by Marian Pugh. The author is grateful for her excellent work. He is indebted to William Bell, E. W. Frees and the referees for helpful comments on earlier versions of this paper. He also wishes to thank the organizers of the Symposium on the Analysis of Statistical Information at the Institute of Statistical Mathematics for the opportunity to present some of these results at this conference and for their very considerate hospitality. Finally, he wishes to acknowledge the support and hospitality of the Institute of Statistical Science of Academia Sinica, Taipei, where the final version of this paper was written.

References

Akaike, H. (1973). Information theory and an extension of the likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Czáki), 267–281, Akadémiai Kiadó, Budapest.

Akaike, H. (1985). Prediction and entropy, *A Celebration of Statistics* (eds. A. C. Atkinson and S. E. Fienberg), Springer, New York.

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

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, 2nd ed., Holden-Day, San Francisco.

Durbin, J. (1960). The fitting of time series models, *Review of the International Institute of Statistics*, **28**, 233–244.

Findley, D. F. (1990). Making difficult model comparisons (submitted for publication).

Findley, D. F. and Wei, C.-Z. (1988). Beyond chi-square: likelihood ratio procedures for comparing non-nested, possibly incorrect regressors, *J. Amer. Statist. Assoc.* (to appear).

Findley, D. F. and Wei, C.-Z. (1991). Bias properties of AIC for possibly incorrect stochastic regression models (in preparation).

Hannan, E. J. and Quinn, B. (1979). The determination of the order of an autoregression, *J. Roy. Statist. Soc. Ser. B*, **41**, 190–195.

Kashyap, R. L. (1980). Inconsistency of the AIC rule for estimating the order of autoregressive models, *IEEE Trans. Automat. Control*, **AC-25**, 996–998.

Levinson, N. (1946). The Wiener RMS (root mean square) error criterion in filter design and prediction, *J. Math. Phys.*, **25**, 261–278.

Poskitt, D. S. (1987). Precision, complexity and Bayesian model determination, *J. Roy. Statist. Soc. Ser. B*, **49**, 199–208.

Raftery, A. E. and Martin, R. D. (1988). Reply, *J. Amer. Statist. Assoc.*, **83**, 1231.

Rissanen, J. (1978). Modelling by shortest data description, *Automatica—J. IFAC*, **14**, 465–471.

Rissanen, J. (1986). Stochastic complexity and modeling, *Ann. Statist.*, **14**, 1080–1100.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117–126.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Ann. Statist.*, **8**, 147–164.

Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45–54 (Correction: ibid. **69**, 494).

Takada, Y. (1982). Admissibility of some variable selection rules in the linear regression model, *J. Japan Statist. Soc.*, **12**, 45–49.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307–333.

Wei, C.-Z. (1991). On predictive least squares principles, *Ann. Statist.* (to appear).

White, H. (1990). *Estimation, Inference and Specification Analysis*, Cambridge University Press, New York.

Woodroofe, M. (1982). On model selection and arc sine laws, *Ann. Statist.*, **10**, 1182–1194.