BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR–93/04

THE OVERFITTING PRINCIPLES SUPPORTING AIC

by

David F. Findley
Statistical Research Division
U. S. Bureau of the Census
Washington, DC 20233–4200

Report issued: April 9, 1993

THE OVERFITTING PRINCIPLES SUPPORTING AIC

DAVID F. FINDLEY
Bureau of the Census
Washington, DC 20233–4200

ABSTRACT

In the context of statistical model estimation and selection, what is "overfit"? What is "overparameterization"? When is a "principle of parsimony" appropriate? Suggestive answers are usually given to such questions, rather than precise definitions and mathematical statistical results. In this article, we investigate some relations that yield asymptotic equality between a variate which is the natural measure of overfit due to parameter estimation and one which is a natural measure of the accuracy loss that occurs when the estimated model is applied to an independent replicate of the data used for estimation. Relations connecting overfit with accuracy loss are what we call *overfitting principles.* The principles we consider yield a theoretical framework in which questions like those posed above can be answered with some precision and with allowance for the possibility that the model family does not contain the true model. One of the relations is shown to be conditionally equivalent to the bias–correction property used by Akaike to motivate the definition of AIC. Our results establishing this principle also provide the first complete verifications of AIC's bias–correction property for general exponential families for i.i.d. data and for invertible Gaussian ARMA time series models.

KEY WORDS: Principle of Parsimony; Minimax Likelihood Principle; Exponential families; ARMA models.

1. INTRODUCTION

Broadly useful methods like Akaike's minimum AIC criterion for model comparison (MAIC) usually rest on one or more deeper theoretical principles of considerable interest. In the case of MAIC, two kinds of principles have been identified. First, Akaike (1973, 1977, 1985) stressed the role of AIC's bias–correction property in the search for the estimated model with the maximal expected loglikelihood (Kullback–Leibler–entropy maximization principle). Linhart and others, see Linhart and Zucchini (1986), have considered extensions of this approach to a variety of expected discrepancy functions different from Kullback–Leibler entropy. Second, Shibata (1980, 1981) discovered a *predictive efficiency* characterization of MAIC in both stochastic and fixed regression contexts. An analog of this predictive efficiency property has been shown to be a useful benchmark for bandwidth selection in nonparametric regression, see Härdle and Marron (1985).

In this note, we focus on the approximation relation which underlies the bias–correction point of view and on a second which is closely related. Both are interpreted as "overfitting principles", because they assert that the natural measure of overfit defined by the criterion optimized for parameter estimation approximates the average loss of fit which occurs whenever the model determined by the estimated parameters is applied to an *independent replicate* of the data set used for estimation. These principles make it possible to give precise formulations and analyses of the fundamental but usually vague concepts of "overfit" and "overparameterization", as well as of the "principle of parsimony". They show how "overfitting" is a general and undesirable phenomenon and that "overparameterization" is a somewhat problematic concept except in special situations involving the comparison of *nested* models. The "principle of parsimony" is similarly limited.

2. The Overfitting Principles and Some Examples to Which They Apply

For the sake of explicitness, our initial discussion will be in terms of maximum likelihood estimation. Let $x_1,...,x_N$ be observed variates whose *log density function* L_N^{true} ($= L_N^{true}(x_1,...,x_N)$) has finite expectation $E_N^{true} \equiv \mathcal{E}\{L_N^{true}\} < \infty$. Let $L_N[\theta]$ ($= L_N[\theta](x_1,...,x_N)$), $\theta \in \Theta$ denote a parametric family of loglikelihood functions which is intended to provide an approximation to L_N^{true}, even if this log density does not belong to the family. The parameter set Θ will be a subset of space $\mathbb{R}^s$ of s–dimensional real column vectors. Thus $\dim\theta = s$. We assume that the parameterization is unique: $L_N[\theta] \neq L_N[\bar{\theta}]$ if $\theta \neq \bar{\theta}$. Further, both $L_N[\theta]$ and the *expected log likelihood function*,

$$E_N[\theta] \equiv \mathcal{E}\{L_N[\theta]\},$$

must be *twice continuously differentiable* on Θ and have *maximizers* over this set, denoted $\hat{\theta}_N$, respectively, θ_N.

The "overfitting principles" referred to earlier are approximation relations of the form

$$L_N[\hat{\theta}_N] - L_N[\theta_N] \doteq E_N[\theta_N] - E_N[\hat{\theta}_N].$$

Such relations will be shown to be naturally interpretable as

(OP) "overfit" $\doteq$ "accuracy loss with independent replicates".

Different measures of approximation define the different principles discussed below.

2.1 Three Examples Classes

Before discussing the various relations, we introduce the three model classes in which all overfitting principles will be verified, sometimes after further restricting the parameter set. The first class is an elementary one, involving linear regression, chosen for its illustrative value.

Example 1. Suppose the observation x_n are *independent*, scalar variates having a mean function $\mu_n = \mathcal{E}x_n$ which can vary with n, but having a *constant variance* v^x and a *constant fourth moment* for $e_n \equiv x_n - \mu_n$. Let z_n be any scalar regressor sequence which satisfies the two conditions $\Sigma_{n=1}^{\infty} z_n^2 = \infty$ and

$$\sup_N N^{-1}\{\sum_{n=1}^{N} \mu_n^2 - (\sum_{n=1}^{N} z_n^2)^{-1}(\sum_{n=1}^{N} \mu_n z_n)^2\} < \infty.$$

The latter holds, for example, if $\sup_n|\mu_n|<\infty$, or if $\mu_n = \beta_0 z_n$ for some β_0. With θ^T denoting transpose of θ, set $\Theta \equiv \{[v\ \beta]^T\colon\ 0 < v < \infty, -\infty < \beta < \infty\}$ and

$$L_N[v,\beta] \equiv -\frac{N}{2}\log 2\pi v - \frac{1}{2v}\Sigma_{n=1}^{N}(x_n - \beta z_n)^2.$$

Then

$$E_N[v,\beta] = -\frac{N}{2}\log 2\pi v - \frac{1}{2v}\{Nv^x + \sum_{n=1}^{N}(\mu_n - \beta z_n)^2\}.$$

The maximizers $\hat{\theta}_N = [\hat{v}_N\ \hat{\beta}_N]^T$ and $\theta_N = [v_N\ \beta_N]^T$ are given by $\hat{\beta}_N \equiv \Sigma_{n=1}^{N} z_n x_n/\Sigma_{n=1}^{N} z_n^2$, $\hat{v}_N \equiv N^{-1}\Sigma_{n=1}^{N}(x_n - \hat{\beta}_N z_n)^2$, $\beta_N \equiv \Sigma_{n=1}^{N} z_n\mu_n/\Sigma_{n=1}^{N} z_n^2$ and $v_N \equiv v^x + N^{-1}\Sigma_{n=1}^{N}(\mu_n - \beta_N z_n)^2$. Since

$$\sum_{n=1}^{N} (\mu_n - \beta_N z_n)^2 = \sum_{n=1}^{N} \mu_n^2 - \Big(\sum_{n=1}^{N} z_n^2\Big)^{-1} \sum_{n=1}^{N} \mu_n z_n,$$

the sequence v_N is *bounded.* This need not be true of the sequence β_N. For example, if $\mu_n = \mu \neq 0$ and if $z_n = n^{-r}$ with $0 < r < 0.5$, then $\beta_N \to \infty$ at the rate N^r (and $v_N \to v^x + \mu^2$) as $N \to \infty$.

Example 2. Consider loglikelihoods for *independent* and *identically distributed* variates x_n being modeled via an s–dimensional *minimal exponential family* of densities (with respect to some measure ν). Thus,

$$L_N[\theta] = - N\kappa[\theta] + \sum_{n=1}^{N} \log b(x_n) + \theta^T \sum_{n=1}^{N} t(x_n),$$

and

$$E_N[\theta] = N\{-\kappa[\theta] + \mathcal{E}\log b(x_n) + \theta^T \mathcal{E} t(x)\},$$

where $\kappa[\theta]$ is an infinitely differentiable and strictly convex function of θ in the interior of the natural parameter set

$$\Theta^* = \{\theta: \int_X e^{\theta^T t(x)} b(x) d\nu < \infty\}.$$

The interior of Θ^*, denoted $\mathrm{Int}\Theta^*$, is the set of $\theta \in \Theta^*$ for which there is an s–dimensional ball of positive radius centered at θ which is entirely contained in Θ^*. See Chapters 7–8 of Barndorff–Nielsen (1978) for information about such families. (*Regular* families, with $\Theta^* = \mathrm{Int}\Theta^*$, include multivariate normal, Poisson, logarithmic

and Hardy–Weinberg distributions, as well as multinomials and negative multinomials. See Barndorff–Nielsen (1970) for complete details.) It follows that $L_N[\theta]$ and $E_N[\theta]$ are infinitely differentiable and strictly concave on $\text{Int}\Theta^*$. So their maximizers, if they are in $\text{Int}\Theta^*$, will be unique. The maximizer of $E_N[\theta]$ does not depend on N, because $E_N[\theta] = NE_1[\theta]$. (This always happens when the x_n are identically distributed, and $L_N[\theta] = \Sigma_{n=1}^N h(x_n)$ for some h(x)). For this example, we assume that $E_N[\theta]$ has a maximizer $\theta_\infty \in \text{Int}\Theta^*$ and that $\mathcal{E}t(x_1)^T t(x_1) < \infty$, conditions which are certainly satisfied if $L_N^{\text{true}} = L_N[\theta_\infty]$ with $\theta_\infty \in \text{Int}\Theta$.

Example 3. Let $x_1,\ldots,x_N$ be successive observations from a covariance stationary time series with mean 0. Suppose we wish to model their autocovariance structure with models which can be defined by spectral density functions $f[\theta](\lambda)$. Set $\gamma_j \equiv \mathcal{E}x_n x_{n-j}$ and

$$\gamma_j[\theta] \equiv \int_{-\pi}^{\pi} f[\theta](\lambda)\cos j\lambda d\lambda \qquad (j = 0, \pm 1, \ldots).$$

Define the N×N matrices $\Gamma_N \equiv [\gamma_{j-k}]_{1\le j,k\le N}$, $\Gamma_N[\theta] \equiv [\gamma_{j-k}[\theta]]_{1\le j,k\le N}$, and the vector $X_N = [x_1 \ \ldots \ x_N]^T$. Then the Gaussian loglikelihood family

$$L_N[\theta] \equiv -\frac{N}{2}\log 2\pi|\Gamma_N[\theta]| - \frac{1}{2} X_N^T \Gamma_N[\theta]^{-1} X_N$$

has the expected loglikelihood function

$$E_N[\theta] = -\frac{N}{2}\log 2\pi|\Gamma_N[\theta]| - \frac{1}{2}\,\text{tr}\Gamma_N[\theta]^{-1}\Gamma_N,$$

where tr denotes trace. Our strongest assertion, (2.3) below, will hold when the time series x_n is Gaussian, Θ is convex and compact, and the $L_N[\theta]$ are loglikelihoods of invertible ARMA models, with $L_N^{true} = L_N[\theta_\infty]$ for some $\theta_\infty \in \text{Int}\Theta$, see Section 6.

In Examples 2 and 3, *explicit* formulas for the maximizers $\hat{\theta}_N$ and θ_N are available only in special cases.

2.2 The Overfitting Principles

For sequences of random variables u_N and v_N, we write $u_N \sim_p v_N$ if $u_N - v_N$ converges to zero in probability as $N \to \infty$ (alternatively, $u_N - v_N \to_p 0$). If $\mathcal{E}\{u_N - v_N\} \to 0$, we write $u_N \sim_{\mathcal{E}} v_N$. Finally, we write $u_N \sim_{mabs} v_N$ for mean absolute convergence, $\mathcal{E}|u_N - v_N| \to 0$. Since $|\mathcal{E}u_N - \mathcal{E}v_N| \le \mathcal{E}|u_N - v_N|$, this condition implies $u_N \sim_{\mathcal{E}} v_N$ as well as $u_N \sim_p v_N$, see Theorem 4.1.4 of Chung (1968, p. 64), for example. This paper is concerned with the interpretation and verification of the two asymptotic relations

$$L_N[\hat{\theta}_N] - L_N[\theta_N] \sim_p E_N[\theta_N] - E_N[\hat{\theta}_N] \tag{2.1}$$

and

$$L_N[\hat{\theta}_N] - L_N[\theta_N] \sim_{\mathcal{E}} E_N[\theta_N] - E_N[\hat{\theta}_N], \tag{2.2}$$

and with some of their consequences.

The relation (2.1) certainly suggests (2.2), but (2.2) is normally much harder to verify, especially when no explicit formula is available for $\hat{\theta}_N$. Section 6 presents the first complete results for this situation. Instead of obtaining (2.2) directly, we shall

verify the stronger result

$$L_N[\hat{\theta}_N] - L_N[\theta_N] \sim_{mabs} E_N[\theta_N] - E_N[\hat{\theta}_N] \ . \tag{2.3}$$

2.3 Interpretation of (2.1) and (2.2)

Let $\bar{x}_1,...,\bar{x}_N$ be an independent replicate of $x_1,...,x_N$. Denote its log density $L_N^{true}(\bar{x}_1,...,\bar{x}_N)$ by $\bar{L}_N^{true}$. Set $\bar{L}_N[\theta] \equiv L_N[\theta](\bar{x}_1,...,\bar{x}_N)$. The expression on the right in (2.1) and (2.2) is non–negative and can be rewritten as a difference of Kullback–Leibler discrepancies (which are quasi–distance measures, see Csiszar (1975)),

$$E_N[\theta_N] - E_N[\hat{\theta}_N] = \{E_N^{true} - E_N[\hat{\theta}_N]\} - \{E_N^{true} - E_N[\theta_N]\}. \tag{2.4}$$

The first expression on the right in (2.4) is the discrepancy between (the distributions defined by) $\bar{L}_N^{true}$ and $\bar{L}_N[\hat{\theta}_N]$. The second is the smallest possible discrepancy between $\bar{L}_N^{true}$ and any $\bar{L}_N[\theta]$, $\theta \in \Theta$, because θ_N maximizes $E_N[\theta]$. By Jensen's inequality, or the inequality of Kullback and Leibler (1951), these quantities are positive unless $\bar{L}_N[\hat{\theta}_N] = \bar{L}_N^{true}$, respectively, $\bar{L}_N^{true}[\theta_N] = \bar{L}_N^{true}$ (with probability one, a qualifying statement we omit henceforth), the latter occurring if and only if L_N^{true} belongs to the parametric family. The variate (2.4) is thus the *excess discrepancy* (above the minimum) arising from the fact that the best approximator, $\bar{L}_N[\theta_N]$ is not known but must be *estimated.* It is an *unavoidable cost of estimation.*

Analogously, $L_N[\theta_N]$ is a best approximation to L_N^{true}, and *maximization* of $L_N[\theta]$ leads to the larger–than–ideal variate $L_N[\hat{\theta}_N]$. (This is the most natural meaning of the statement "$\hat{\theta}_N$ is overfitting".) Under (2.1), its excess $L_N[\hat{\theta}_N] - L_N[\theta_N]$ above $L_N[\theta_N]$ approximates the excess Kullback–Leibler discrepancy (2.4)

associated with $\hat{\theta}_N$. Under this approximation therefore, $L_N[\hat{\theta}_N] - L_N[\theta_N]$ can be naturally reinterpreted as a *cost of overfitting.* The second asymptotic relation (2.2) makes a similar statement in terms of averages. Thus, if we call the variate on the left in (2.1)–(2.3) the *overfit* and the variate on the right the *loss of accuracy,* these relations provide instances of the statement (OP) given at the beginning of this subsection. This is why they are called *overfitting principles.* As the counterexample of Section 7 demonstrates, they are not universally valid.

In familiar situations, both sides of (2.1) have a limiting distribution which a second–degree Taylor expansion argument shows to be the distribution of a linear combination of $s = \dim\theta$ independent chi–square variates, each with 1 degree of freedom, having nonnegative coefficients,

$$L_N[\hat{\theta}_N] - L[\theta_N],\ E_N[\theta_N] - E_N[\hat{\theta}_N] \rightarrow_{\text{dist.}} \frac{1}{2}\sum_{i=1}^{s}\beta_i^2\chi_i^2(1)\ . \qquad (2.5)$$

If the true log density is contained in the parametric family, then usually $\beta_i^2 = 1$ for i=1,...,s. In general, therefore, overfit is a random quantity which varies from realization to realization of the observed sample. In the special situation in which $\beta_i^2 = 1$ for all i, overfit depends, asymptotically, only on the number of parameters estimated.

Finally, we note that it follows from (2.1) that if θ_N and $\bar{\theta}_N$ are distinct maximizers of $E_N[\theta]$, then $L_N[\theta_N] \sim_p L_N[\bar{\theta}_N]$.

2.4 Connections with AIC

From the discussion above, it is clear that the quantity $E_N[\hat{\theta}_N]$ would be a natural performance measure to apply to competing model families for purposes of

model selection if enough information about its value could be obtained from the observed quantity $L_N[\hat{\theta}_N]$. To have an *asymptotically unbiased* estimator, it is clear from the identity

$$E_N[\hat{\theta}_N] = L_N[\hat{\theta}_N] + \{E_N[\hat{\theta}_N] - L_N[\hat{\theta}_N]\} \tag{2.6}$$

that an estimate of $\mathcal{E}\{E_N[\hat{\theta}_N] - L_N[\hat{\theta}_N]\}$ is what is needed. It follows from $\mathcal{E}\{L_N[\theta_N]\} = E_N[\theta_N]$ and

$$\begin{aligned} E_N[\hat{\theta}_N] - L_N[\hat{\theta}_N] = &\{E_N[\hat{\theta}_N] - E_N[\theta_N]\} \\ &+ \{E_N[\theta_N] - L_N[\theta_N]\} + \{L_N[\theta_N] - L_N[\hat{\theta}_N]\} \end{aligned} \tag{2.7}$$

that the overfitting principle (2.2) is equivalent to

$$E_N[\hat{\theta}_N] - L_N[\hat{\theta}_N] \sim_{\mathcal{E}} -2\{L_N[\hat{\theta}_N] - L_N[\theta_N]\} \ . \tag{2.8}$$

The conditions we utilize in Theorem 3.2 below to derive (2.2) and (2.3) also yield, via (2.5), that the bias approximator on the right in (2.8) satisfies

$$2\{L_N[\hat{\theta}_N] - L_N[\theta_N]\} \sim_{\mathcal{E}} \sum_{i=1}^{s} \beta_i^2 \ . \tag{2.9}$$

Given (2.9), it follows from (2.8) that (2.2) *is equivalent to*

$$E_N[\hat{\theta}_N] \sim_{\mathcal{E}} L_N[\hat{\theta}_N] - \sum_{i=1}^{s} \beta_i^2 \ . \tag{2.10}$$

This is noteworthy because, when the true log density belongs to the model family, resulting in $\beta_i^2 = 1$, (2.10) can be rewritten as the *bias relation* used in Akaike (1973) to motivate the *definition of AIC*, which, in our notation, is

$$-2E_N[\hat{\theta}_N] \quad \sim_{\mathcal{E}} \quad AIC_N \equiv -2L_N[\hat{\theta}_N] + 2\dim\theta \ . \tag{2.11}$$

That is, in this situation, (2.2) *and* (2.11) *are equivalent.*

It is clear from (2.6) that $AIC_N[\hat{\theta}_N]$ does not estimate $-2E_N[\hat{\theta}_N]$ in any other probabilistic sense, because the mean zero term $E_N[\theta_N] - L_N[\theta_N]$ on the right in (2.7) will generally be of order $N^{1/2}$ in probability and will dominate the other terms, which are bounded in probability, by (2.5): consider, for example, the case in which $x_1,\ldots,x_N$ are i.i.d., $L_N[\theta] = \Sigma_{n=1}^N \log f[\theta](x_n)$, and $\log f[\theta_\infty](x)$ has finite variance v_∞. Then $N^{-1/2}\{E_N[\theta_\infty] - L_N[\theta_\infty]\}$ converges in distribution to $\mathcal{N}(0,v_\infty)$ and so is bounded in probability. Hence, $-2E_N[\hat{\theta}_N] - AIC_N[\hat{\theta}_N]$ is of order $N^{1/2}$ in probability.

The fact that $AIC_N[\hat{\theta}_N]$ approximates $-2E_N[\hat{\theta}_N]$ only in the mean sense was originally made clear in Shimizu (1978), where the first complete published statement and proof of (2.1) can be found, for the case of overparameterized Gaussian autoregressions estimated by least squares. Shimizu does not mention the interpretation of (2.1) as an overfitting principle given in the preceding subsection.

At least in the case in which L_N^{true} belongs to the parametric family, (2.1) and (2.2) are now part of the lore of AIC. The relation (2.1) is suggested in Figure 4.2 of Sakamoto, Ishiguro and Kitagawa (1986, p. 68), which is presented in the course of their informal derivation of AIC for i.i.d. observations. (Our development of (2.10) is similar in outline to their derivation of (2.11)). It will be seen in the next section that second–order Taylor expansions easily suggest (2.1) and (2.2). What

have been lacking are complete verifications of (2.2) and (2.9), and a general formulation and derivation of (2.1).

Remark. Shibata (1989) discusses estimators of the term $\Sigma_{i=1}^{s} \beta_i^2$ in (2.10) given independent observations, including the first such estimator, proposed by Takeuchi (1976), for a special case. He also considers the situation in which maximum *penalized* likelihood estimators are used. More generally, Ishiguro et al. (1991) successfully utilize a bootstrap estimator of the bias term $\mathcal{E}\{E_N[\hat{\theta}_N] - L_N[\hat{\theta}_N]\}$ for estimators $\hat{\theta}_N$ which need not satisfy (2.1).

3. THEOREMS AFFIRMING (2.1)–(2.3)

The theorems of this section do not require that $L_N[\theta]$ be a loglikelihood function. They thereby cover alternative situations considered by Linhard and Zucchini (1986). For example, sums of squared errors from linear or nonlinear least squares regressions can satisfy the conditions we impose. The basic requirements are that $L_N[\theta] = L_N[\theta](x_1,...,x_N)$ and its expected value function $E_N[\theta] \equiv \mathcal{E}\{L_N[\theta]\}$ be twice continuously differentiable on the convex set spanned by Θ, that Θ have nonempty interior in $\mathbb{R}^s$, and that $L_N[\theta] \neq L_N[\bar{\theta}]$ with positive probability whenever $\theta \neq \bar{\theta}$ (*identifiability*). We use $L_N'[\theta]$ to denote the gradient vector $\partial L_N[\theta]/\partial\theta$ and $L_N''[\theta]$ to denote the Hessian matrix $\partial^2 L_N[\theta]/\partial\theta\partial\theta^T$. Similarly for $E_N'[\theta]$ and $E_N''[\theta]$. Our derivations focus on the analysis of the difference of two second–degree Taylor expansions, of $L_N[\theta_N] - L_N[\hat{\theta}_N]$ about $\hat{\theta}_N$, and of $E[\hat{\theta}] - E[\theta_N]$ about θ_N, elaborated by the insertion of a nonsingular normalizing matrix B_N,

$$\{L_N[\theta_N] - L_N[\hat{\theta}_N]\} - \{E_N[\hat{\theta}_N] - E_N[\theta_N]\} =$$

$$(\theta_N - \hat{\theta}_N)^T L_N'[\hat{\theta}_N] - (\hat{\theta}_N - \theta_N)^T E_N'[\theta_N]$$

$$+ \tfrac{1}{2}\{B_N(\hat{\theta}_N - \theta_N)\}^T\{B_N^{-T}(L_N''[\tilde{\theta}_N] - E_N''[\bar{\theta}_N])B_N^{-1}\}\{B_N(\hat{\theta}_N - \theta_N)\}, \tag{3.1}$$

where $\tilde{\theta}_N$ and $\bar{\theta}_N$ have the form $\alpha_N\theta_N + (1-\alpha_N)\hat{\theta}_N$ with α_N a random variable satisfying $0 < \alpha_N < 1$, and where B_N^{-T} denotes $(B_N^T)^{-1}$. Often B_N is $N^{1/2}$ times the identity matrix of order $\dim\theta$, but, for Example 1, $B_N \equiv$ $\mathrm{diag}(N^{1/2}, (\Sigma_{n=1}^N z_n^2)^{1/2})$ is appropriate. We now present an easy and general result for (2.1) and a more restricted theorem affirming (2.3). The assumptions have been chosen to accommodate some situations in which the sequence θ_N does not converge, as in the special case discussed for Example 1.

Theorem 3.1. Suppose that the sequences $\hat{\theta}_N = \hat{\theta}_N(x_1,\ldots,x_N)$ *and* θ_N *satisfy* (P1) *below. Assume also that there is a sequence of nonsingular matrices* B_N *such that the eigenvalues of* $B_N^T B_N$ *tend to* ∞ *as* $N \to \infty$, *and such that conditions* (P2)–(P3) *hold:*

(P1) (i) $E_N'[\theta_N] = 0$ *for all* $N \geq N_0$, *for some* N_0.

(ii) *There is a* $K > 0$ *such that if* $(\hat{\theta}_N - \theta_N)^T(\hat{\theta}_N - \theta_N) < K$, *then* $L_N'[\hat{\theta}_N] = 0$.

(P2) *The sequence* $B_N(\hat{\theta}_N - \theta_N)$ *is bounded in probability.*

(P3) (i) $B_N^{-T}\{E_N''[\theta_N\} - L_N''[\theta_N]\}B_N^{-1} \to_p 0$.

Also, for any sequence $\theta_N^* = \alpha_N\theta_N + (1 - \alpha_N)\hat{\theta}_N$ *with* $\alpha_N = \alpha_N(x_1,\ldots,x_n)$ *satisfying* $0 < \alpha_N < 1$, *we have*

(ii) $$B_N^{-T}\{L_N''[\theta_N^*] - L_N''[\theta_N]\}B_N^{-1} \to_p 0,$$

and

(iii) $$B_N^{-T}\{E_N''[\theta_N^*] - E_N''[\theta_N]\}B_N^{-1} \to_p 0.$$

Then (2.1) *holds.*

Proof. We must show that the terms on the right in (3.1) converge to 0 in probability. For the second derivative expression, this follows immediately from (P2) and(P3), because $B_N^{-T}\{L_N''[\tilde{\theta}_N] - E_N''[\bar{\theta}_N]\}B_N^{-1}$ is a linear combination of expressions of the kind considered in (P3). The $E_N'[\theta_N]$ term is ultimately 0, by (P1i). Finally, because of the assumption on the eigenvalues of $B_N^T B_N$, (P2) implies $\hat{\theta}_N - \theta_N \to_p 0$. Hence, by (P1ii), the probability that $(\hat{\theta}_N - \theta_N)^T L_N'[\theta_N]$ is 0 tends to one. This completes the proof.

To obtain (2.3) from (2.1), and (2.9) from (2.5), we will use the following standard result, see Chung (1968, p. 452)).

Lemma 3.1. If the sequence of variates z_N *satisfies* $z_N \to_{\text{dist.}} z$ *for some* z, *and if* $\sup_N \mathcal{E}|z_N|^p < \infty$ *for some* $p > 1$, *then* $\mathcal{E}|z_N| \to \mathcal{E}|z|$, *and* $\mathcal{E}z_N \to \mathcal{E}z$.

For $p > 0$ and for random z, consider the p–norm $\|z\|_p \equiv \{\mathcal{E}|z|^p\}^{1/p}$. To verify the Lemma's moment boundedness assumption, we will utilize a simple variant of Hölder's inequality: if the positive numbers p_1, p_2, p_3 are such that $p_1^{-1} + p_2^{-1} + p_3^{-1} = p^{-1}$, then $\|uvw\|_p \leq \|u\|_{p_1}\|v\|_{p_2}\|w\|_{p_3}$, see Hardy, Littlewood and Pólya (1952, p. 44), for example. In particular, for $\alpha > 0$,

$$\|uvw\|_{1+\alpha/(2+\alpha)} \leq \|u\|_{4(1+\alpha)}\|v\|_2\|w\|_{4(1+\alpha)} . \tag{3.4}$$

We partition the integers $j = 1,...,s$ into the (possibly empty) subset J of indices j such that $\partial L_N[\hat{\theta}_N]/\partial\theta_j = 0$ holds with probability 1 for $N \geq N_0$ for some N_0, and its complementary subset, denoted $\bar{J}$ (possibly empty).

Theorem 3.2. Let the hypotheses of Theorem 3.1 be satisfied and also (E1) – (E3) *below.*

(E1). *For some* $\alpha>0$, $\sup_N\|(\hat{\theta}_N - \theta_N)^T B_N^T B_N(\hat{\theta}_N - \theta_N)\|_{2(1+\alpha)} < \infty$.

(E2). *For any* j *in the complementary subset* $\bar{J}$ *of coordinate indices defined above,* $B_{N,jj} = N^{1/2}$, *and* $B_{N,ij} = 0$ *if* $i\neq j$, $i = 1,...,s$. *Further,*

$$\sup_N \|N^{-1}\partial L_N[\hat{\theta}_N]/\partial\theta_j\|_2 < \infty .$$

(E3). *For every sequence* θ_N^* *of the sort considered in* (P3),

$$\text{(i)}\quad \sup_N \|\mathrm{tr}(B_N^{-T}E_N''[\theta_N^*]B_N^{-1})\|_2 < \infty ,$$

$$\text{(ii)}\quad \sup_N \|\mathrm{tr}(B_N^{-T}L_N''[\theta_N^*]B_N^{-1})\|_2 < \infty .$$

Then (2.3) *is satisfied. Further, if* $L_N[\hat{\theta}_N] - L_N[\theta_N]$ *converges in distribution* to a *variate* z *with a finite mean, then*

$$\lim_{N\to\infty} \mathcal{E}\{L_N[\hat{\theta}_N] - L_N[\theta_N]\} = \mathcal{E}z = \lim_{N\to\infty} \mathcal{E}\{E_N[\theta_N] - E_N[\hat{\theta}_N]\}. \quad (3.5)$$

Proof. By the Lemma and (2.1), it suffices to establish the $1 + \alpha/(2+\alpha)$–norm boundedness of the Taylor expansion terms on the right in (3.1). For the second degree terms, this follows from (E1) and (E3) via (3.4). The term $(\hat{\theta}_N - \theta_N)^T E_N'[\theta_N]$ is eventually zero by (P1i).

It remains to consider terms involving $\partial L_N[\hat{\theta}_N]/\partial\theta_j$ with $j\in\bar{J}$. Let A(N) denote the event $\{(\hat{\theta}_N - \theta_N)^T(\hat{\theta}_N - \theta_N) \geq K\}$, pr(A(N)) its probability, and $1_{A(N)}$ the variate which is 1 if this event is true and zero otherwise. By (P2ii), for each $j\in\bar{J}$,

$$(\hat{\theta}_{N,j} - \theta_{N,j})\partial L_N[\hat{\theta}_N]/\partial\theta_j = \{N^{1/2}(\hat{\theta}_{N,j} - \theta_{N,j})\}\{N^{-1}\partial L_N[\hat{\theta}_N]/\partial\theta_j\}\{N^{1/2}1_{A(N)}\} .$$

Therefore, by (3.4) and (E1) – (E2), the boundedness result we are after can be obtained by verifying

$$\sup_N \{N^{2(1+\alpha)}\mathrm{pr}(A(N))\}^{1/4(1+\alpha)} \; (= \sup_N \|N^{1/2}1_{A(N)}\|_{4(1+\alpha)}) < \infty .$$

This follows easily from Chebyshev's inequality (Chung (1976, p. 46)):

$$\sup_N N^{2(1+\alpha)}\mathrm{pr}(A(N)) \leq K^{-2(1+\alpha)}\sup_N N^{2(1+\alpha)}\mathcal{E}((\hat{\theta}_N - \theta_N)^T(\hat{\theta}_N - \theta_N))^{2(1+\alpha)}$$

$$= K^{-2(1+\alpha)} \sup_N \mathcal{E}(N(\hat{\theta}_N - \theta_N)^T(\hat{\theta}_N - \theta_N))^{2(1+\alpha)} < \infty ,$$

the finiteness resulting from (E1).

There are numerous results available in the literature for verifying (P2) and (P3). Some references are given in Section 5 below. The situation is quite different with (E1) – (E3). Precise results are given in Section 6 to establish their validity in cases of Examples 1–3.

4. MINIMAX LIKELIHOOD PRINCIPLES, PARSIMONY, AND ANTIPARSIMONY

The left hand side of (2.1) is not an observable quantity because θ_N is unknown. We can only observe differences of $L_N[\hat{\theta}_N]$–values from competing models. Under the conditions described in this section, such differences approximate the difference in overfitting costs of the estimated models. As in Section 2, we assume that θ_N and $\hat{\theta}_N$ are maximizers of $E_N[\theta]$ and $L_N[\theta]$, respectively.

Suppose we have two competing families for approximating L_N^{true}, namely $L_N^{(i)}[\theta^{(i)}]$, $\theta^{(i)} \in \Theta^{(i)} \subseteq \mathbb{R}^{s(i)}$, $i = 1,2$. The relations

$$L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}] \sim_p E_N^{(2)}[\hat{\theta}_N^{(2)}] - E_N^{(1)}[\hat{\theta}_N^{(1)}] \tag{4.1}$$

and

$$L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}] \sim_{\mathcal{E}} E_N^{(2)}[\hat{\theta}_N^{(2)}] - E_N^{(1)}[\hat{\theta}_N^{(1)}] \tag{4.2}$$

follow immediately from (2.1) and (4.3) – (4.4) below, respectively, (2.2) and (4.4):

$$L_N^{(1)}[\theta_N^{(1)}] - L_N^{(2)}[\theta_N^{(2)}] \to_p 0, \tag{4.3}$$

$$E_N^{(1)}[\theta_N^{(1)}] - E_N^{(2)}[\theta_N^{(2)}] \rightarrow 0. \qquad (4.4).$$

Note that (4.3) and (4.4) both hold if

$$L_N^{(1)}[\theta_N^{(1)}] = L_N^{(2)}[\theta_N^{(2)}], \qquad N \geq N_0 \ . \qquad (4.5).$$

If L_N^{true} belongs to both families, then $L_N^{(i)}[\theta_N^{(i)}] = L_N^{true}$, i = 1,2 as we explained in Section 2.2, and (4.5) is satisfied.

The left–hand side of (4.4) is the difference of the Kullback–Leibler discrepancies of the best models from the true model, $E_N^{true} - E_N^{(i)}[\theta_N^{(i)}]$, i = 1,2. Thus (4.4) requires these best models to be asymptotically equidistant from the truth. Usually, when (4.4) fails, $E_N^{(1)}[\theta_N^{(1)}] - E_N^{(2)}[\theta_N^{(2)}]$ has order N. Then the costs of overfit, $E_N[\theta_N^{(1)}] - E_N[\hat{\theta}_N^{(1)}]$, i = 1,2, which are bounded (in probability) by (2.5), will be of negligible importance for model selection, for large enough N.

The condition (4.3) requires that the best models be asymptotically coincident. In many situations where (4.3) holds, both sides of (4.1) have a limiting distribution which is the distribution of a linear combination of independent chi–square variates with 1 d.f.,

$$L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}], \ E_N^{(2)}[\hat{\theta}_N^{(2)}] - E_N^{(1)}[\hat{\theta}_N^{(1)}] \rightarrow_{\text{dist.}} \sum_{j=1}^{m} \alpha_j \chi_j^2(1). \qquad (4.6)$$

In (4.6), both positive and negative coefficients α_j can occur unless the models are nested, see Theorem 4.3 of Vuong (1989, p. 313) for the i.i.d. case and Proposition 7.3 of Findley and Wei (1989) for the case of stochastic regression models.

4.1 Minimax Likelihood Principles

To help interpret (4.1), we note that it has the following consequence. Suppose that for some $\delta > 0$, the events

$$B_N(\delta) = \{L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}] > \delta\}$$

have a probability which is ultimately positive, $\lim_{N\to\infty} \mathrm{pr}(B_N(\delta)) > 0$. Then for any $0 < \delta^* < \delta$, it follows from (4.1) that the *conditional* probability under $B_N(\delta)$ that $E_N^{(2)}[\hat{\theta}_N^{(2)}] - E_N^{(1)}[\hat{\theta}_N^{(1)}] > \delta^*$ holds will converge to 1,

$$\lim_{N\to\infty} \mathrm{pr}(\, E_N^{(2)}[\hat{\theta}_N^{(2)}] - E_N^{(1)}[\hat{\theta}_N^{(1)}] > \delta^* \mid B_N(\delta)) = 1,$$

because this probability is bounded below by

$$\mathrm{pr}(|\{E_N^{(2)}[\hat{\theta}^{(2)}] - E_N^{(1)}[\hat{\theta}^{(1)}]\} - \{L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}]\}| < \delta\text{-}\delta^* \mid B_N(\delta)).$$

In this sense in particular, the value of $L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}]$ is indicative of the value of $\{E_N^{\mathrm{true}} - E_N^{(1)}[\hat{\theta}_N^{(1)}]\} - \{E_N^{\mathrm{true}} - E_N^{(2)}[\hat{\theta}_N^{(2)}]\}$. Thus, in the case of loglikelihood functions, (4.1) embodies a *minimax likelihood principle*: among models which are asymptotically equivalent in the sense that (4.3) and (4.4) hold, the model with *smaller* maximum likelihood is to be preferred, because it has a lower cost of overfit, in the sense discussed in Section 2.3, by approximately the amount $|L_N^{(1)}[\hat{\theta}_N^{(1)}] - [L_N^{(2)}[\hat{\theta}_N^{(2)}]|$.

The relation (4.2) expresses an analogous *mean minimax loglikelihood principle* concerning *mean overfit.*

4.2 Nested Models: Overparameterization and the Parsimony Principle

We say that $L_N^{(2)}[\theta^{(2)}]$, $\theta^{(2)} \in \Theta^{(2)}$ is ***nested*** in $L_N^{(1)}[\theta^{(1)}]$, $\theta^{(1)} \in \Theta^{(1)}$ if, for each $\theta^{(2)} \in \Theta^{(2)}$, there is a $\theta^{(1)} \in \Theta^{(1)}$ such that $L_N^{(2)}[\theta^{(2)}] = L_N^{(1)}[\theta^{(1)}]$ holds for all N. Because of our identifiability requirement, this $\theta^{(1)}$ is unique, and the function g defined by $g(\theta^{(2)}) = \theta^{(1)}$ has the property that $g(\theta^{(2)}) \neq g(\bar{\theta}^{(2)})$ if $\theta^{(2)} \neq \bar{\theta}^{(2)}$. Assuming that $\dim\theta^{(2)} < \dim\theta^{(1)}$, we say that the $\theta^{(1)}$–family is *overparameterized relative to the smaller* $\theta^{(2)}$*–family* if the E_N–functions ultimately have the same maximum value,

$$E_N^{(1)}[\theta_N^{(1)}] = E_N^{(2)}[\theta_N^{(2)}] \quad (N \geq N_0) \; . \tag{4.7}$$

In other words, the best models in both classes become equidistant from L_N^{true} in the Kullback–Leibler sense. Assuming that the maximizer $\theta_N^{(1)}$ of $E_N^{(1)}[\theta^{(1)}]$ is unique when $N \geq N_0$, it follows from (4.7) that $g(\theta_N^{(2)}) = \theta_N^{(1)}$ for $N \geq N_0$, and hence that (4.5) holds. Then, if (2.1) and (2.2) hold for the two families, so do (4.1) and (4.2). Since $L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}]$ is nonnegative, the limiting distribution in (4.6) will be positive with probability one. Therefore, the *mean minimax principle* will prefer the nested $\hat{\theta}_N^{(2)}$–model when N is sufficiently large, and the *minimax principle* will also prefer this more parsimonious model with probability approaching one, as $N \to \infty$. In these senses, *overparameterization is undesirable.*

These two asymptotic arguments in favor of the model with fewer parameters when (4.7) holds constitute precise, mathematical derivations of two *principles of* (parameter) *parsimony*, derivations which *avoid* the assumption that L_N^{true} is contained in the nested family. Somewhat surprisingly perhaps, no similarly general principle holds for non–nested models comparisons, as we now explain.

4.3 Non–Nested Models: Parsimony and Antiparsimony

With non–nested models satisfying (4.5) which are not close to the correct model, it can happen that the mean of the distribution of the right hand side of (4.6) is such that the mean minimax loglikelihood principle favors the family with *more* estimated parameters, and so is *antiparsimonious*, see the second example of Findley (1991). On the other hand, if both families contain L_N^{true} and if, say, $\dim\theta^{(2)} < \dim\theta^{(1)}$, then (4.6) usually takes the form

$$2\{L_N^{(1)}[\hat\theta_N^{(1)}] - L_N^{(2)}[\hat\theta_N^{(2)}]\},\quad 2\{E_N^{(2)}[\hat\theta_N^{(2)}] - E_N^{(1)}[\hat\theta_N^{(1)}]\} \rightarrow_{\text{dist.}} \Delta(m,d), \tag{4.8}$$

with $\Delta(m,d) \equiv \chi^2(m+d) - \chi^2(d)$, a difference of *independent* chi–square variates with degrees of freedom d+m and m, where $d \equiv \dim\theta^{(1)} - \dim\theta^{(2)}$, and $m \equiv \dim\theta^{(2)} - \dim\theta^{(1,2)}$, with $\dim\theta^{(1,2)}$ denoting the largest dimension of a family nested in both the $\theta^{(1)}$– and $\theta^{(2)}$–families, see Proposition 7.2 of Findley and Wei (1989), for example. Then, under the assumptions of Theorem 3.2, we obtain

$$E_N^{(2)}[\hat\theta_N^{(2)}] - E_N^{(1)}[\hat\theta_N^{(1)}] \sim_{\mathcal{E}} \tfrac{1}{2}\, d, \tag{4.9}$$

which shows that, in this case, the *mean minimax loglikelihood principle selects the more parsimonious $\theta^{(2)}$–model* when N is sufficiently large. This is a weaker *principle of parsimony for nonnested models which overparameterize the true model.* However, since

$$\lim_{N\rightarrow\infty} \text{pr}(E_N^{(2)}[\hat\theta_N^{(2)}] - E_N^{(1)}[\hat\theta_N^{(1)}] > 0) = \text{pr}(\Delta(m,d) > 0)$$

is less than one, it follows that there is always a non–zero probability that $\Delta(m,d)$ is negative. For this event, the minimax likelihood principle selects the *less parsimonious* $\theta^{(1)}$– model, because it has less overfit, for large enough N. Table 1 below gives these *probabilities of antiparsimony* for a range of values of m and d.

Table 1

The probability $\mathrm{pr}(\Delta(m,d) < 0)$ *that the m.l.e. model with* d *additional parameters is less overfitting asymptotically than the more parsimonious model, when the dimension of the intersection of the two competing loglikelihood families is* m *less than the dimension of the more parsimonious family. It is assumed that the true log density belongs to both families.*

m\d	1	2	6	12	18	∞
1	.29	.18	.03	.00	.00	0
2	.35	.25	.06	.01	.00	0
6	.42	.35	.14	.03	.01	0
12	.44	.39	.21	.07	.02	0
18	.45	.41	.25	.10	.04	0
∞	.50	.50	.50	.50	.50	0

Let us now consider the performance of MAIC under (4.8). Assuming (4.1) and (4.2) also hold, the difference of AIC values,

$$\mathrm{AIC}_N^{(1,2)} \equiv -2\{L_N^{(1)}[\hat{\theta}_N^{(1)}] - L_N^{(2)}[\hat{\theta}_N^{(2)}]\} + 2d$$

is an asymptotically unbiased estimate of $-2\{E_N^{(1)}[\hat{\theta}_N^{(1)}] - E^{(2)}[\hat{\theta}_N^{(2)}]\}$, but does not estimate this quantity in any stronger sense, being, in fact, a consistent estimator of

$$-2\{E_N^{(2)}[\hat{\theta}_N^{(2)}] - E_N^{(1)}[\hat{\theta}_N^{(1)}]\} + 2d,$$

by (4.1). The derivation of AIC given in Subsection 2.3 makes it clear that the minimum AIC procedure is focused on selecting the model with smaller *mean* overfitting cost, which is the more parsimonious $\theta^{(2)}$–model. How well does it succeed? Under (4.8), the limiting probability of selection of the $\theta^{(2)}$–model by MAIC is given by

$$\lim_{N\to\infty} \mathrm{pr}(\mathrm{AIC}_N^{(1,2)} > 0) = \mathrm{pr}(\Delta(m,d) < 2d) .$$

Table 2 gives these values for several choices of m and d. The values associated with m=0 are for the nested case. The tabled values show that MAIC achieves its goal best when $\dim\theta^{(1)}-\dim\theta^{(2)}$ is large relative to the parameter excess in the intersection of the families.

Table 2

Asymptotic probability that the minimum AIC procedure chooses the more parsimonious model under (4.8) for a range of m and d values.

m\d	1	2	6	12	18	∞
0	.84	.87	.94	.98	.99	1.00
1	.74	.81	.92	.98	.99	1.00
2	.68	.77	.90	.97	.99	1.00
6	.59	.67	.85	.95	.98	1.00
12	.56	.62	.79	.92	.97	1.00
18	.55	.59	.75	.89	.95	1.00
∞	.50	.50	.50	.50	.50	

We note, finally, that since $\Delta(m,d) < 0$ excludes $\Delta(m,d) > 2$, the limiting probability is zero that MAIC selects the *less* parsimonious model when this is the model preferred by minimax loglikelihood principle.

Remark. The derivation of (4.8) given in Findley and Wei (1989) is for stochastic regression models, but, using the notion of orthogonal parameters, see Cox and Reid (1987), it is easy to see how to carry out generalizations to other situations, such as that of Theorem 3.3 of Vuong (1989).

5. VERIFICATION OF (P1) – (P3) OF THEOREM 3.1

Consider first the situation (appropriate for Examples 2 and 3) in which $\hat{\theta}_N$ and θ_N converge into a compact set in the interior of Θ, as happens, for example, if $\theta_N \longrightarrow \theta_\infty \in \mathrm{Int}\Theta$ and (P2) holds. Then (P1) is satisfied, and (P3) will follow from Uniform Laws of Large Numbers and their analogues, ensuring uniform convergence in probability of $B_N^{-T}\{L_N''[\theta] - E_N''[\theta]\}B_N^{-1}$, and from uniform equicontinuity of $B_N^{-T}E_N''[\theta]B_N^{-1}$, on compact sets: see Pötscher and Prucha (1991a) for very general results and examples related to dynamic nonlinear regression and also covering the i.i.d. case. Of course, in the case of our Example 2, where $N^{-1}L_N''[\theta]$ and $N^{-1}E_N''[\theta]$ coincide with the continuous matrix function $-\kappa''[\theta]$, such results are automatic.

The remaining condition (P2) is satisfied if $B_N(\hat{\theta}_N - \theta_N)$ has a limiting distribution. Results establishing asymptotic normality are given in Berk (1972) for exponential families and in Pötscher and Prucha (1991b) for dynamic regression models as well as i.i.d. models, for which a quite general discussion has also been given by Bryant (1991).

Finally, let us consider Example 1, where θ_N can be unbounded. Here (P1) is satisfied because $L_N'[\hat{\theta}_N] = 0 = E_N'[\theta_N]$. For (P2) – (P3), consider the formulas

$$(\sum_{n=1}^{N} z_n^2)^{1/2}(\hat{\beta}_N - \beta_N) = (\sum_{n=1}^{N} z_n^2)^{-1/2} \sum_{n=1}^{N} z_n e_n , \tag{5.1}$$

$$N^{1/2}(\hat{v}_N - v_N) = N^{-1/2} \sum_{n=1}^{N} (e_n^2 - v^x) - N^{-1/2} (\sum_{n=1}^{N} z_n^2)(\hat{\beta}_N - \beta_N)^2$$

$$+ 2N^{-1/2} \sum_{n=1}^{N} (\mu_n - \beta_N z_n) e_n \,, \tag{5.2}$$

and, as representatives of the second derivative formulas,

$$2N^{-1} \frac{\partial^2}{\partial v^2} L_N[v,\beta] = v^{-2} - 2v^{-3}\{\hat{v}_N + N^{-1}(\beta - \hat{\beta}_N)^2 \sum_{n=1}^{N} z_n^2\} \,, \tag{5.3}$$

$$2N^{-1} \frac{\partial^2}{\partial v^2} E_N[v,\beta] = v^{-2} - 2v^{-3}\{v_N + N^{-1}(\beta - \beta_N)^2 \sum_{n=1}^{N} z_n^2\} \,. \tag{5.4}$$

From (5.1), we see that $(\Sigma_{n=1}^{N} z_n^2)^{-1/2}(\hat{\beta}_N - \beta_N)$ is bounded in probability because its mean square is bounded (equal to v^x). Similarly for the first term on the right in (5.2), because e_n has a finite fourth moment, and for the third term, since $N^{-1}\Sigma_{n=1}^{N}(\mu_n - \beta_N z_n)^2$ is bounded. Consequently, (P2) holds. The way in which (P3) is also satisfied is now clear from expressions like (5.3) and (5.4), because the terms explicitly involving β_N^* tend to zero in probability, as does $v_N^* - v_N$.

6. VERIFICATION OF (E1) – (E3) OF THEOREM 3.2

Not surprisingly, we have to impose more restrictions on the model families to obtain (2.3). We will often need the parameter set Θ to be compact, usually a compact subset of a noncompact "natural" parameter set. Also, when no explicit formula is available for $\hat{\theta}_N$, we will assume that the true model is in the family, at some $\theta_\infty \in \text{Int}\Theta$.

6.1 When $\hat{\theta}_N$ is a Linear Least Squares Estimate.

We focus mainly on Example 1. Here the deepest problem is the difficulty of verifying (E3) caused by the negative powers of v_N^* in $L_N''[v_N^*, \beta_N^*]$ and $E_N''[v_N^*, \beta_N^*]$, see (5.3) – (5.4).

Note that since $v_N \geq v^x$, only the event $\hat{v}_N \leq v_N^* \leq v_N$ is problematic. Hence it suffices to establish

$$\sup_{N \geq N(p)} \mathcal{E}|\hat{v}_N|^{-p} < \infty \quad (p = 1,2,3,\ldots) . \tag{6.1}$$

In general, this requires strong assumptions. We consider two possibilities.

6.1.1. Unrestricted Θ.

Assume that x_n is Gaussian and that $\mu_n = \beta_0 z_n$. In this case, $\hat{v}_N$ has a $\chi^2(N-1)$ distribution, and (6.1) holds for all $p \geq 1$ with $N(p) = 2p+1$, see Lemma 2 of Sriram and Bose (1988) for a general result. Clearly (E3) follows easily in this case, as do (E1) ($\bar{J}$ is void) and (E2).

Findley and Wei (1989) established the analogues of (6.1) and (2.3) for Gaussian vector autoregressions. Bhansali and Papangelou (1992) improved upon their results in the case of univariate autoregressions. Findley and Wei (1993) give a very general result for non–Gaussian vector autoregressions.

6.1.2. Restricted v: $\Theta = \{[v\ \beta]^T\colon\ v_0 \leq v < \infty, -\infty < \beta < \infty\}$ for some $0 < v_0 < v^x$.

Assume that $\sup_N \mathcal{E}|e_n|^{8(1+\alpha)} < \infty$, and that the regressors z_n satisfy

$$\lim_{N\to\infty} \Big(\sum_{n=1}^{N} z_n^2\Big)^{-4(1+\alpha)} \sum_{n=1}^{N} |z_n|^{8(1+\alpha)} = 0, \tag{6.2}$$

and

$$\sup_N N^{-1} \sum_{n=1}^{N} |\mu_n - \beta_N z_n|^{4(1+\alpha)} \equiv D_\infty < \infty, \tag{6.3}$$

for some $\alpha > 0$. It is easy to check that (6.2) is satisfied if $\mu_n = \mu \neq 0$ and $z_n = n^{-r}$, $0 \leq r < 0.25(1+\alpha)^{-1}$. So is (6.3), because then $N^{-1}\beta_N^{4(1+\alpha)}\Sigma_{n=1}^{N}|z_n|^{4(1+\alpha)}$ has order $N^{-1}N^{4(1+\alpha)r}N^{1-4(1+\alpha)r} = 1$.

The formula given earlier for the maximizing value v_N in the case of unrestricted Θ remains valid, because $v_N \geq v_x > v_0$. But now the $\hat{v}_N$ maximizing $L_N[v, \hat{\beta}_N]$ has the formula

$$\hat{v}_N = \max\{v_0, N^{-1}\Sigma_{n=1}^{N}(x_N - \hat{\beta}_N z_n)^2\} . \tag{6.4}$$

If $|\hat{v}_N| < 0.5(v^x - v_0)$, then $\hat{v}_N > v^x$ and $L_N'[\hat{v}_N, \hat{\beta}_N] = 0$. Hence (P1ii) holds.

The condition (6.2) is exactly what is required to obtain the convergence of $\mathcal{E}|N^{1/2}(\hat{\beta}_N - \beta)|^{8(1+\alpha)}$ to $\mathcal{E}|Z|^{8(1+\alpha)}$, with $Z \sim \mathcal{N}(0, v^x)$, see Corollary 2 of Chow and Teicher (1988, p. 410). Hence these moments are bounded.

On the other hand, a standard moment inequality for sums of independent variates, see Lemma III.3.1 of Ibragimov and Has'minskii (1981, p. 186), asserts that D_∞ in (6.3) is proportional to an upper bound of the absolute $4(1+\alpha)$–moments of the final term on the right of the identity (5.2) for $N^{-1/2}(N^{-1}\Sigma_{n=1}^{N}(x_n - \hat{\beta}_N z_n)^2 - v_N)$. That is, (6.3) yields

$$\sup_N \|N^{-1/2}\Sigma_{n=1}^{N}(\mu_n - \beta_N z_n)e_n\|_{4(1+\alpha)} < \infty .$$

Under (6.4), since $v_N > v_0$, we have

$$|N^{1/2}(\hat{v}_N - v_N)| \leq |N^{1/2}(\Sigma_{n=1}^{N}(x_n - \hat{\beta}_N z_n)^2 - v_N|.$$

Thus (6.2) and (6.3) together imply, via (5.2), that $\sup_N \|N^{1/2}(\hat{v}_N - v_N)\|_{4(1+\alpha)} < \infty$. Therefore (E1) holds. As is clear from (5.3) – (5.4), these results, and the fact that $\hat{v}_N^{-1} \leq v_0^{-1}$, are enough to establish (E2) and (E3).

6.2 No Explicit Formula for $\hat{\theta}_N$: I.I.D. Case (Example 2)

The only source of difficulty in verifying the conditions of Theorem 3.2 for general minimal exponential family models is the verification of (E1). We begin with a general result for (E1). Suppose that the x_n, n = 1,2,... are independent and identically distributed, and that $f[\theta](x)$, $\theta \in \Theta$ is a family of density functions (with respect to a σ–finite measure ν on the sample space X) containing the true density of the x_n, $f^{true}(x) = f[\theta_\infty](x)$ with $\theta^\infty \in \text{Int}\Theta$. We need a set of conditions involving only Θ and the $f[\theta]$ which yields (E1). The main such result in the literature is Theorem III. 3.2 of Ibragimov and Has′minskii (1981). Theorem 6.1 below is a conveniently stated special case, some of whose hypotheses we now describe in advance.

We consider only the case where Θ is *compact* set in $\mathbb{R}^s$ with nonempty interior and *without isolated points* (every point on the boundary is a limit of points in the interior). As always, $f[\theta](x)$ and $f[\bar{\theta}](x)$ are required to define distinct distributions if $\theta \neq \bar{\theta}$. Also, $f[\theta](x)$ must be differentiable at every $\theta \in \Theta$ for ν–almost all x, and $f^{1/2}[\theta](x)\partial \log f[\theta](x)/\partial\theta$ must be *ν–square integrable.* Further, at each $\theta \in \Theta$, the "information" matrix

$$I[\theta] \equiv \int_X \{\partial \log f[\theta](x)/\partial\theta\}\{\partial \log f[\theta](x)/\partial\theta\}^T f[\theta](x)d\nu$$

must be *nonsingular.* Then, under the assumptions of Theorem 6.1 below, $I[\theta]$ is a continuous function of θ, see Theorem I.7.1 of Ibragimov and Has′minskii (1981). As a consequence, its eigenvalues are bounded and bounded away from zero on Θ. Finally, if $s > 1$, we also require that for some $r>s$,

$$\sup_\theta \int_X \sum_{i=1}^{s} |\partial \log f[\theta]/\partial\theta_i|^r f[\theta]d\nu < \infty . \tag{6.5}$$

Recall that a family $g[\theta](x)$, $\theta \in \Theta$ is called L^2_ν*–continuous on* Θ if

$$\lim_{\bar{\theta} \to \theta} \int_X |g(\bar{\theta}) - g(\theta)|^2 d\nu = 0 \tag{6.6}$$

holds for all $\theta \in \Theta$. ($\bar{\theta}$ is restricted to lie in Θ).

Theorem 6.1 (Ibragimov and Has'minskii). *Suppose that the conditions stated above apply and, also, that* $f^{1/2}[\theta]$ *and* $\partial f^{1/2}[\theta]/\partial\theta_i$, $i = 1,...,s$ *are* L^2_ν*–continuous. Then the maximum likelihood estimates* $\hat{\theta}_N$ *satisfy*

$$N^{1/2}(\hat{\theta}_N - \theta_\infty) \to_{\text{dist.}} \mathcal{N}(0, I[\theta_\infty]^{-1}) , \tag{6.7}$$

and, for every $p \geq 1$,

$$\sup_N \mathcal{E}|N(\hat{\theta}_N - \theta_\infty)^T(\hat{\theta}_N - \theta_\infty)|^{p/2} < \infty . \tag{6.8}$$

We utilize this result in the proof of the following theorem, which verifies the bias–correction property of AIC for exponential family variates.

Theorem 6.2. Let x_n, $n = 1,...,$ *be* i.i.d. *variates with an exponential family distribution which is specified in terms of a minimal sufficient statistic* $t(x_n)$ *of dimension* s *by means of a density* $f[\theta_\infty](x_n) = b(x_n)\exp\{\theta_\infty^T t(x_n) - \kappa[\theta]\}$ *(with respect to a* σ*–finite measure* ν*), with* θ_∞ *in the interior of the natural parameter space* Θ^*. *Let* Θ *denote any compact set in* $\text{Int}\Theta^*$ *containing* θ_∞ *in its interior, and having no isolated points. Then* (2.3) *holds, for arbitrary* $\alpha > 0$, *as does* (2.11) *for the maximizer* $\hat{\theta}_N$ *of* $L_N[\theta] \equiv N^{-1} \Sigma_{n=1}^N \log f[\theta](x_n)$ *over* Θ.

Proof. We will verify the conditions of Theorem 3.2 with $B_N = N^{-1/2}I$. Because θ_∞ defines the true density, it is the maximizer of $E_N[\theta]$, and since $\theta_\infty \in \text{Int}\Theta^*$, the condition (P1i) holds with

N_0=1. We have $N^{-1}L_N'[\theta] = N^{-1}\,\Sigma_{n=1}^{N}\, t(x_n) - \kappa'[\theta]$. Thus the validity of (E2) follows from the fact that $\kappa'[\theta]$ is continuous on Θ, so its entries are bounded on this set. The remaining conditions follow from $N^{-1}L_N''[\theta] = N^{-1}E_N''[\theta] = -\kappa''[\theta]$ and from verifying the hypotheses of Theorem 6.1. This will establish (E1) and also (2.5) with $\beta_i^2 = 1$, $i \le i \le s$, since, for example,

$$L_N[\hat{\theta}_N] - L_N[\theta_\infty] = (N/2)(\hat{\theta}_N - \theta_\infty)^T \kappa''[\theta_N^*](\hat{\theta}_N - \theta_\infty),$$

and we will have $N^{1/2}(\hat{\theta}_N - \theta_\infty) \longrightarrow_{\text{dist}} \mathcal{N}(0, \kappa''[\theta_\infty]^{-1})$, from (6.7).

It remains, therefore, to verify the L_ν^2–continuity of $f^{1/2}[\theta](x)$ and its gradient, and the condition (6.5). The required continuity of $f^{1/2}[\theta](x)$ is apparent from

$$\begin{aligned}\int_X \{f^{1/2}[\bar{\theta}](x) - f^{1/2}[\theta](x)\}^2 d\nu &= \int\{f[\bar{\theta}] + f[\theta] - 2f^{1/2}[\bar{\theta}]f^{1/2}[\theta]\}d\nu \\ &= 2 - 2\exp\{-(\kappa[\bar{\theta}] + \kappa[\theta])/2\}\int_X \exp\{(\bar{\theta}+\theta)^T t(x)/2\}d\nu \\ &= 2 - 2\exp\{\kappa[(\bar{\theta}+\theta)/2] - (\kappa[\bar{\theta}] + \kappa[\theta])/2\}, \qquad (6.9)\end{aligned}$$

because of the continuity of $\kappa[\theta]$. Now consider the derivatives

$$\begin{aligned}\partial f^{1/2}[\theta](x)/\partial\theta_i &= f^{1/2}[\theta]\partial \log f[\theta]/\partial\theta_i \\ &= f^{1/2}[\theta]t_i(x) - f^{1/2}[\theta]\kappa_i'[\theta], \qquad (6.10)\end{aligned}$$

for i = 1, ..., s. The L_ν^2–continuity of the final term in (6.10) follows from that of $f^{1/2}[\theta]$ and the continuity of $\kappa_i'[\theta]$. The L_ν^2–continuity of $\partial f^{1/2}[\theta](x)/\partial\theta_i$ is therefore an immediate consequence of the formula

$$\int_X t_i^2(x)\{f^{1/2}[\bar{\theta}] - f^{1/2}[\theta]\}^2 d\nu = \kappa''_{ii}[\bar{\theta}] + \kappa''_{ii}[\theta]$$

$$- 2\,\kappa''_{ii}[(\theta+\theta)/2]\exp\{\kappa[(\bar{\theta}+\theta)/2] - (\kappa[\theta] + \kappa[\theta])/2\},$$

which is derived by a calculation analogous to (6.9), using the fact that

$$\kappa''_{ii}[\theta] = \int_X t_i^2(x) f[\theta](x) d\nu\,.$$

It remains to establish (6.5) for $\partial \log f[\theta]/\partial\theta = t(x) - \kappa'[\theta]$. Since $\kappa'[\theta]$ is bounded on compact sets, by continuity, this can be accomplished by verifying

$$\sup_\Theta \int_X |t(x)^T t(x)|^q f[\theta](x) d\nu < \infty \qquad (6.11)$$

for some integer q such that $2q > s$. For this, we only have to recall that 2q–th moments can be written as polynomials in cumulants of order 2q and less, which, for t(x), are the partial derivatives of $\kappa[\theta]$ of order 2q and less. Since these are bounded on Θ, (6.11) follows.

6.3 No Explicit Formula For $\hat{\theta}_N$: Gaussian ARMA Case (Example 3)

Maliukevicius (1988) was able to build on the results of Chapter III Ibragimov and Has′minskii (1981) to obtain an analogue of Theorem 6.1 for Gaussian time series. The theorem below is a simply–stated special case of his Theorems 2.1 and 2.2.

Let x_n be a Gaussian, mean zero, invertible ARMA time series with spectral density $f^{true}(\lambda)$. Let $f[\theta](\lambda)$, $\theta \in \Theta$, be a parametric family of invertible ARMA spectral densities such that Θ is convex and compact, and $f(\theta) \neq f[\theta]$ if $\theta \neq \theta$. Then $f[\theta](\lambda)$ is differentiable on Θ and it is

known, see Poskitt and Tremayne (1981, pp. 976–977), and also Hosoya and Taniguchi (1982, p. 138), that the matrix

$$V[\theta] = \int_{-\pi}^{\pi} \{\frac{\partial}{\partial\theta} \log f[\theta](\lambda)\}^T \{\frac{\partial}{\partial\theta} \log f[\theta](\lambda)\} d\lambda$$

is nonsingular for all $\theta \in \Theta$. This is an essential condition for

Theorem 6.2. (Maliukevicius) *If* $f^{true}(\lambda) = f[\theta_\infty](\lambda)$ *for some* $\theta_\infty \in \text{Int}\Theta$, *then under the assumptions made above, we have*

$$N^{1/2}(\hat{\theta}_N - \theta_\infty) \longrightarrow_{\text{dist.}} \mathcal{N}(0, V[\theta_\infty]^{-1}), \tag{6.11}$$

and

$$\sup_N \mathcal{E}|N(\hat{\theta}_N - \theta_\infty)^T(\hat{\theta}_N - \theta_\infty)|^{p/2} < \infty\ , \tag{6.12}$$

for all $p \geq 1$.

To verify the conditions of Theorem 3.2 under the assumptions of this theorem, we first note that $N^{1/2}(\theta_N - \theta_\infty) \longrightarrow 0$ holds, as explained in Remark 2.1 of Findley (1985), so that (P1), (E1), and (2.5) with $\beta_i^2 = 1$, $1 \leq i \leq s$ follow from (6.11) – (6.12). Then, utilizing (6.12), the arguments used to establish (3.8) – (3.10) of Findley (1985) immediately yield (P3) and (E2) – (E3). Hence, by Theorem 3.2, *the relations* (2.3) *and* (2.11) *are valid for these models.*

Remark. In Findley (1985), this conclusion was obtained by *assuming* the validity of (E1) and by making a stronger assumption than (P1ii), essentially that $\hat{\theta}_N \in \text{Int}\Theta$ with probability one. The latter condition is too strong: as the situation analyzed in Section 6.1.2 illustrates, for each N, $\hat{\theta}_N$ can be expected to be on the boundary with positive probability, because the innovation variance parameter is not naturally bounded away from zero.

7. A COUNTEREXAMPLE TO (2.1) AND (2.2)

For a sequence e_n of independent Gaussian variates each with mean zero and variance 1, let x_n be the random walk process defined by $x_0 \equiv 0$ and by $x_n \equiv \Sigma_{t=1}^{n} e_t$ for $n \geq 1$. With $-\infty < \theta < \infty$, consider

$$L_N[\theta] \equiv -\frac{1}{2} \sum_{n=1}^{N} (x_n - \theta x_{n-1})^2 .$$

Since $\mathcal{E}x_n x_{n-1} = \mathcal{E}x_{n-1}^2 = n-1$ if $n \geq 1$, it follows that the expected–value function $E_N[\theta]$ has the formula

$$E_N[\theta] = -\frac{1}{2} \sum_{n=1}^{N} \{\mathcal{E}x_n^2 + (\theta^2 - 2\theta)\mathcal{E}x_{n-1}^2\} = -\frac{1}{2}\{N + (\theta-1)^2 N(N-1)/2\} .$$

Hence its maximizer is $\theta_N = 1$. For $\hat{\theta}_N = \Sigma_{n=1}^{N} x_n x_{n-1} / \Sigma_{n=1}^{N} x_{n-1}^2$, we have

$$L_N[\hat{\theta}_N] - L_N[1] = \left(\sum_{n=1}^{N} x_{n-1}^2\right)(\hat{\theta}_N - 1)^2$$

and

$$E_N[1] - E_N[\hat{\theta}_N] = (N(N-1)/2)(\hat{\theta}_N - 1)^2 .$$

The variates on the right are known to have different limiting distributions, see Example 3 of Lai and Wei (1982) and Corollary 3.13 of Chan and Wei (1988), for example. Therefore (2.1) fails to hold. Concerning (2.2), David Dickey has provided the author with the following values, obtained from well–tested approximations to the asymptotic distributions, values we confirmed with Monte Carlo estimates of the expressions on the left,

$$\lim_{N \to \infty} \mathcal{E}\{L_N[\hat{\theta}_N] - L_N[1]\} \doteq 1.1,$$

$$\lim_{N \to \infty} \mathcal{E}\{E_N[1] - E_N[\hat{\theta}_N]\} \doteq 6.2.$$

Thus, not surprisingly, (2.2) also fails.

ACKNOWLEDGEMENT

Some of the results presented here were obtained while the author visited the Department of Mathematics of the University of Lancaster as a Senior Research Fellow, sponsored by the Science and Engineering Research Council of the United Kingdom. He is grateful for the support and for the hospitality he received during this visit, especially from Granville Tunnicliffe–Wilson.

References

Akaike, H. (1985). Prediction and entropy, *A Celebration of Statistics*, eds. A. C. Atkinson and S. E. Fienberg, Springer–Verlag: New York, 1–24.

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Adademia Kiado: Budapest, 267–281. Reprinted in Kotz and Johnson (1992).

Barndorff–Nielsen, O. (1970). *Exponential Families, Exact Theory*, Mathematics Institute Various Publications Series, No. 19. Aarhus University.

Barndorff–Nielsen, O. (1978). *Information and Exponential Families in Statistics.* Wiley: New York.

Berk, R. H. (1972). Consistency and asymptotic normality of mle's for exponential models. *Annals of Mathematical Statistics*, 43, 193–204.

Bhansali, R. J. and Papangelou, F. (1991). "Convergence of moments of least–squares estimators for the coefficients of an autoregressive process of unknown order. *Annals of Statistics*, 19, 1155–1162.

Bryant, P. G. (1991). Large–sample results for optimization–based clustering methods. *Journal of Classification*, 8, 31–44.

Chan, N. H. and Wei, C.–Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics*, 16, 367–401.

Chow, S.–C. and H. Teicher (1988). *Probability Theory, 2nd Edition.* Springer–Verlag: New York.

Chung, Kai–Lai (1968). *A Course in Probability Theory.* Harcourt, Brace and World: New York.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society* Series B, 49, 1–39.

Csiszar, I. (1975). I–divergence geometry of probability distributions and minimization problems, *Annals of Probability*, 3, 146–158.

Dunsmuir, W. and Hannan, E. J. (1976). Vector linear time series models. *Advances in Applied Probability*, 8, 339–364.

Findley, D. F. (1985). On the unbiasedness property of AIC for exact or approximating linear time series models, *J. Time Series Analysis*, 6, 229–252.

Findley, D. F. and Wei, C.–Z. (1989). Beyond chi–square: likelihood ratio procedures for comparing non–nested, possibly incorrect regressors. *Statistical Research Division Report RR 89/08.* Bureau of the Census.

Findley, D. F. (1991). Counterexamples to parsimony and BIC. *Annals of the Institute of Statistical Mathematics*, 43, 505–514.

Findley, D. F. and Wei, C.–Z. (1993). Vector autoregressions, AIC, and the negative moments of sample variance matrices. (in preparation)

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, 13, 1465–1481.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation.* Springer–Verlag: New York.

Ishiguro, M., Morita, K. I. and Ishiguro, M. (1991). Applications of an estimator–free information criterion (WIC) to aperture synthesis imaging. *In Radio Interferometry: Theory, Techniques and Applications*, eds. T. J. Cornwell and R. A. Perley, 243–248. International Astronomical Union Coll. 131, Conference Series, 19, Astronomical Society of the Pacific: San Francisco.

Kendall, M. G. and Stuart, A. (1968). *The Advanced Theory of Statistics*, 3, *2nd Edition.* Hafner: New York.

Kotz, S. and Johnson, N. L. (1992). *Breakthroughs in Statistics, Vol. I.* Springer–Verlag: New York.

Kullback, S. and Leibler R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

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

Linhart, H. and Zucchini, W. (1986). Model Selection. Wiley: New York.

Maliukevicius, R. (1988). "Maximum Likelihood Estimation of the Spectral Density Parameter." Lithuanian Mathematical Journal, 28, 353–364.

Poskitt, D. S. and Tremayne, A. R. (1981). "An Approach to Testing Linear Time Series Models." Annals of Statistics, 9, 974–986.

Pötscher, B. M. and Prucha, I. R. (1991a). "Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models, part I: Consistency and Approximation Concepts." Econometric Reviews, 10, 125–216.

Pötscher, B. M. and Prucha, I. R. (1991b). "Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models, part II: Asymptotic Normality." Econometric Revieux, 10, 253–325.

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

Shibata, R. (1980). "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process." Annals of Statistics, 8, 147–164.

Shibata, R. (1981). "An Optimal Selection of Regression Variables." Biometrika 68, 45–54.

Shibata, R. (1989). "Statistical Aspects of Model Selection," in From Data to Model, ed. J. C. Willems. Springer–Verlag: Berlin, 215–240.

Shimizu, R. (1978). "Entropy Maximization Principle and Selection of the Order of an Autoregressive Gaussian Process." Annals of the Institute of Statistical Mathematics, 30, 263–270.

Sriram, T. N. and Bose, A. (1988). "Sequential Shrinkage Estimation in the General Linear Model." Sequential Analysis, 7, 149–163.

Takeuchi, K. (1976). "Distribution of Information Statistics and a Criterion for Model Fitting." Suri–Kagaku (Mathematical Sciences) 153, 12–18 (in Japanese).