The Annals of Statistics
1982, Vol. 10, No. 4, 1182–1194

ON MODEL SELECTION AND THE ARC SINE LAWS[1]

BY MICHAEL WOODROOFE

University of Michigan

Generalizations of the arc sine laws are shown to provide insight into the operating characteristics of certain techniques for selecting models to fit a given data set, when the available models are nested. As a corollary, one sees that a popular technique may be expected to include about one superfluous parameter, even if the sample size is large.

1. Introduction. There are several techniques which may be used to select an appropriate model from a class of available models to fit a given data set. In particular, Mallows's (1964, 1973) C_p criterion and Akaike's (1974) entropy maximization criterion have been recommended for use in model selection. Assuming that one of the available models is correct, one may inquire about such operating characteristics as the probability of finding the smallest correct model and the distribution of the number of superfluous parameters in the model selected. Here these operating characteristics are studied in the special case that the models are nested—as, for example, in polynomial regression and moving average models for time series. In this case there is a natural relation between the operating characteristics and certain generalizations of the arc sine law, as described by Feller (1966, Chapter 12), for example. Briefly, the selection techniques choose the model for which a criterion is maximum; and the generalized arc sine laws determine the distribution of the index for which sums of i.i.d. random variables attain a maximum. So, if the model selection criteria form sums of i.i.d. random variables, the arc sine laws may be used to determine the distribution of the index of the model selected.

In Section 2 the relation between the operating characteristics of Akaike's technique and the generalized arc sine laws is indicated in the simplest case—when the data are independent, normally distributed random variables with unknown means and unit variances. Then the generalized arc sine laws are reviewed in Section 3, and applied to the normal case in Section 4. A natural extension then yields the operating characteristics for Mallows' C_p in Section 5. In Sections 6 and 7, the simple normal example is shown to provide an asymptotic distribution for the number of superfluous parameters for models with a well-behaved likelihood function and a large sample size. As a corollary, it is noted that Akaike's technique is inconsistent in large samples. The same is true of Mallows' C_p, but not of Schwarz's (1978) Bayesian criterion. These remarks are detailed in Section 8.

There are several related articles. The formulation of the problem as one of selecting from multiple hypotheses is similar to that in Anderson's (1962) determination of the degree to use in polynomial regression. Anderson developed optimal procedures. The emphasis here is on the properties of suboptimal, though closely related, procedures. That Akaike's technique is inconsistent in large samples was shown by Shibata (1976) for autoregressive processes and more generally by Hinkley (1976) in an unpublished manuscript. Much of the present Section 6 was anticipated in the latter. Recently, Shibata (1980, 1981) has studied model selection techniques under a different limiting operation. Under this operation, Akaike's technique and Mallows's C_p are asymptotically efficient.

2. Akaike's Criterion. This technique starts with a large model which is assumed to be correct, but possibly redundant, and eliminates parameters which appear to be

Received November 1981; revised May 1982.

[1] Research supported by the National Science Foundation under MCS-8101897.

AMS 1980 *subject classifications.* 62F99; 62J05.

Key words and phrases. Akaike's criterion, Asymptotic distributions, Mallows C_p, Random walks.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.*
www.jstor.org

superfluous. Thus, let $X = (X_1, \cdots, X_n)'$ denote a random column vector having a density $f(\cdot; \theta)$, where $\theta = (\theta_1, \cdots, \theta_k)'$ is a vector of unknown parameters taking values in an open subset $\Omega \subset R^k$; for $j = 0, \cdots, k$, let

$$\Omega_j = \{\theta \in \Omega: \theta_i = 0, i = j + 1, \cdots, k\};$$

and let $H_j: \theta \in \Omega_j$ be the assertion that the smaller model, with Ω replaced by Ω_j, contains the true distribution of X. Thus, the models are assumed to be nested, as in polynomial regression or moving average models for time series. Next, let

$$\Lambda_j = \sup_{\theta \in \Omega_j} \log f(X; \theta)$$

be the maximum value of the log likelihood, assuming Ω_j, $j = 0, \cdots, k$. Then Akaike's technique selects the model for which

$$\mathrm{AIC}(j) = \Lambda_j - j = \max.$$

Now let θ^0 denote the true value of the parameter. That is, suppose that X has density $f(\cdot; \theta^0)$. Then Akaike's technique selects the model for which

$$\mathrm{AIC}^*(j) = \Lambda_j^* - j = \max,$$

where

$$\Lambda_j^* = \sup_{\theta \in \Omega_j} \log f(X; \theta) - \log f(X; \theta^0)$$

is the log likelihood ratio statistic for testing $\theta = \theta^0$ vs. $\theta \in \Omega_j$, $j = 0, \cdots, k$. Thus, the index of the model selected is

$$J_k = \min[j: \mathrm{AIC}(j) = \max_{0 \le i \le k} \mathrm{AIC}(i)]; \tag{1}$$

and AIC may be replaced by AIC*.

To see the relation between Akaike's Criterion and the generalized arc sine laws, consider the simple special case in which $n = k$, $X_1, \cdots, X_k$ are independently, normally distributed random variables with unknown means $\theta_1, \cdots, \theta_k$ and unit variances, and $\theta^0 = (0, \cdots, 0)'$. Then,

$$\mathrm{AIC}^*(j) = \frac{1}{2} \sum_{i=1}^{j} X_i^2 - j = \sum_{i=1}^{j} \left(\frac{1}{2} X_i^2 - 1\right) = S_j, \quad \text{say},$$

for $j = 0, \cdots, k$. Observe that $S_1, \cdots, S_k$ form an initial segment of a random walk with negative drift, since $E(\frac{1}{2} X_i^2 - 1) = -\frac{1}{2}$, $i = 1, \cdots, k$. Thus,

$$J_k = \min\{j: S_j = \max_{1 \le i \le k} S_i\}; \tag{2}$$

and J_k is the number of superfluous parameters included, since $\theta^0 = 0$. The distribution of random variables of the form J_k have been studied extensively in the context of general random walks $S_1, S_2, \cdots$. Some of the relevant results are reviewed in the next section.

3. The generalized arc sine laws. Let $Y_1, Y_2, \cdots$ be any sequence of i.i.d. random variables, and let S_j, $j \ge 0$, denote the associated random walk, $S_0 = 0$ and $S_j = Y_1 + \cdots + Y_j$, $j \ge 1$. Further, let $p_0 = q_0 = 1$,

$$p_j = P\{S_1 > 0, \cdots, S_j > 0\}$$

and

$$q_j = P\{S_1 \le 0, \cdots, S_j \le 0\}, \quad j \ge 1.$$

Then

$$P\{J_k = j\} = p_j q_{k-j}, \quad \text{for } 0 \le j \le k, k \ge 1, \tag{3}$$

where J_k, $k \ge 1$, are defined by (2). The quantities p_j, $j \ge 0$, and q_j, $j \ge 0$, may be determined from the generating functions

$$P(s) = \sum_{j=0}^{\infty} p_j s^j \quad \text{and} \quad Q(s) = \sum_{j=0}^{\infty} q_j s^j, \quad 0 < s < 1, \tag{4}$$

in view of the following, remarkable identities. Let

$$a_n = P\{S_n > 0\}, \quad n \geq 1;$$

then

$$P(s) = \exp\left\{\sum_{n=1}^{\infty} \frac{1}{n} a_n s^n\right\}, \quad 0 < s < 1, \tag{5a}$$

and

$$Q(s) = \exp\left\{\sum_{n=1}^{\infty} \frac{1}{n} (1 - a_n) s^n\right\}, \quad 0 < s < 1. \tag{5b}$$

For example, if Y_1 has a continuous, symmetric distribution, then $a_n = ½$ for all $n \geq 1$, so that $P(s) = Q(s) = 1/\sqrt{(1-s)}$, $0 < s < 1$, and $P\{J_k = j\} = (-1)^k \binom{-½}{j}\binom{-½}{k-j}$ for $0 \leq j \leq k$ and $k \geq 1$. That is, J_k has the classical, discrete arc sine distribution as described by Feller (1968, Chapter 3), for example. The results of this paragraph are taken directly from Feller (1966, Chapter 12).

The mean and variance of J_k may be computed easily from (3). In fact, $E(J_k) = \sum_{j=0}^{k} j p_j q_{k-j}$ is the coefficient of s^{k-1} in the expansion of $P'(s)Q(s)$; and the latter is easily seen to be $a_1 + \cdots + a_k$. So,

$$E(J_k) = a_1 + \cdots + a_k, \quad k \geq 1; \tag{6}$$

and

$$D(J_k) = \sum_{j=1}^{k} j a_j - \sum\sum_{1 \leq i < j \leq k} a_i a_j, \quad k \geq 1,$$

by a similar argument.

There is natural interest in the distribution of J_k when k is large; and, if $E(Y_1) < 0$, then the latter is easily determined. In fact, if $E(Y_1) < 0$, then the series $\sum_{n=1}^{\infty} n^{-1} a_n$ is convergent and

$$\lim_{k\to\infty} q_k = \lim_{s\uparrow 1}\{(1-s)Q(s)\} = \exp\left(-\sum_{n=1}^{\infty} \frac{1}{n} a_n\right) = q_\infty, \quad \text{say;} \tag{7}$$

so,

$$P\{J_k = j\} \to q_\infty p_j,$$

as $k \to \infty$ for all $j \geq 0$.

4. Normal case. Now reconsider the simple normal case in which $X_1, \cdots, X_k$ are independent normally distributed random variables with means $\theta_1, \cdots, \theta_k$ and unit variances. If $\theta_i = 0$ for all $i \leq k$, then $\mathrm{AIC}^*(j) = Y_1 + \cdots + Y_j$ for $1 \leq j \leq k$ with $Y_i = ½ X_i^2 - 1$ for $1 \leq i \leq k$; and the generalized arc sine laws may be applied directly to find the distribution of J_k, with

$$a_j = P(\chi_j^2 > 2j), \quad 1 \leq j \leq k.$$

For example, the limiting distribution of J_k as $k \to \infty$ is $\lim P_0\{J_k = j\} = q_\infty p_j$ for $j \geq 0$, where p_j, $j \geq 0$, are determined from (5) and q_∞ is as in (7).

Table 1 lists the exact distribution of J_k for $k = 5$, 10 and the limiting distribution as $k \to \infty$. Observe that the probabilities approach their limits quite quickly, but that the convergence of $E(J_k)$ is slower. Observe also that the limiting distribution assigns slightly more than 1% of its mass to integers $j > 10$. Two of the numbers in Tables 1 and 2 are of special interest. If $\theta = 0$ and k is large, then the probability of correctly determining that $\theta = 0$ is approximately 0.712, but the expected number of superfluous parameters included is approximately 0.946.

The exact computations for the special case $\theta = 0$ provide bounds for the general case. To see how, let

TABLE 1
The Distribution of J_k for $k = 5, 10, \infty$.

j	k = 5	k = 10	k = ∞
0	.736	.718	.712
1	.117	.113	.112
2	.061	.058	.057
3	.038	.035	.035
4	.027	.023	.023
5	.022	.016	.016
6		.012	.011
7		.0088	.0083
8		.0067	.0061
9		.0054	.0046
10		.0047	.0035
> 10			.0115
$E(J_k)$	.571	.791	.946

The computations were done on an Apple II microcomputer, using formula (24.4.6) *of Abramowitz and Stegun* (1970) *to compute the Chi squared distribution function. $E(J_k)$ was computed from* (6).

$$\gamma_k(j;\theta) = P_\theta(J_k > j), \tag{8}$$

for $j = 0, \cdots, k$, $k \geq 1$, and $\theta \in R^k$. If $\theta \in \Omega_r - \Omega_{r-1}$, where $1 \leq r < k$, then $\gamma_k(r + j; \theta)$ is the probability that the criterion includes more than j superfluous parameters in the model for $j = 0, \cdots, k - r - 1$.

THEOREM 1. *For $1 \leq r < k$ and $0 \leq j < k - r$,*

$$\sup_{\theta \in \Omega_r - \Omega_{r-1}} \gamma_k(r + j; \theta) = \gamma_{k-r}(j; 0).$$

PROOF. Let $S_0 = 0$ and $S_j = Y_1 + \cdots + Y_j$ for $1 \leq j \leq k$. Then $J_k > r + j$ iff $S_{r+i} > \max(S_0, \cdots, S_{r+j})$ for some $i \leq k - r$. If $\theta \in \Omega_r - \Omega_{r-1}$, then

$$\begin{aligned} P_\theta(J_k > r + j) &= P_\theta\{S_{r+i} > \max(S_0, \cdots, S_{r+j}), \exists\, i \leq k - r\} \\ &\leq P_\theta\{S_{r+i} - S_r > \max(S_r - S_r, \cdots, S_{r+j} - S_r), \exists i \leq k - r\} \\ &= P_0\{S_i > \max(S_0, \cdots, S_j), \exists i \leq k - r\} = \gamma_{k-r}(j; 0). \end{aligned} \tag{9}$$

Moreover, the difference between the first and second lines in (9) is at most

$$P_\theta\{\max(0, S_1, \cdots, S_{r+j}) > \max(S_r, \cdots, S_{r+j}), \exists\, j \leq k - r\} \leq P_\theta\{S_r < \max(0, \cdots, S_r)\}$$

which tends to zero as $\theta_r \to \infty$.

5. Mallows's criterion. Now consider a linear model

$$X = M\beta + \varepsilon,$$

where M is an $n \times k$ matrix of full rank $k < n$ and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)'$ has the n-variate normal distribution with mean vector 0 and covariance matrix $\sigma^2 I_n$. Here the unknown parameters are $\beta \in R^k$ and $\sigma^2 > 0$; and the nested models are

$$\Omega_j = \{\beta \in R^k\colon \beta_i = 0,\ i = j + 1, \cdots, k\}, \quad 0 \leq j \leq k.$$

Let SSE_j denote the error sum of squares when the model $H_j\colon \beta \in \Omega_j$, $0 < \sigma^2 < \infty$ is fit; and let $\hat{\sigma}^2 = \mathrm{SSE}_k/(n - k)$ denote the unbiased estimator of σ^2 when all k regression parameters are fit. Then

$$C_p = \hat{\sigma}^{-2} \cdot \mathrm{SSE}_p + (2p - n)$$

has been suggested as a criterion for judging the adequacy of the models H_p, $0 \le p \le k$. See Mallows (1964, 1973) and, for example, Daniel and Wood (1980, pages 86-90). The distribution of the index

$$\mathscr{J}_k = \min\{p\colon C_p = \min_{1 \le j \le k} C_j\}$$

may be found by a natural extension of the techniques of Sections 3 and 4.

Let $L_j \subset R^k$ be the linear subspace spanned by the first j columns of M, $1 \le j \le k$; and let $e_1, \cdots, e_n$ be an orthonormal basis for R^n for which $e_1, \cdots, e_j$ is an orthonormal basis for L_j for $1 \le j \le k$. Then

$$\mathrm{SSE}_j = \textstyle\sum_{i=j+1}^n Z_i^2, \quad 1 \le j \le k,$$

where

$$Z_i = e_i' Y, \quad 1 \le i \le n,$$

are independent normally distributed random variables with means $\theta_i = e_i' \mathrm{M}\beta$ for $1 \le i \le k$, and $e_i' M\beta = 0$ for $k < i \le n$, and common variance σ^2. See, for example, Lehmann (1959, Section 7.2). It follows that

$$C_0 - C_p = \hat{\sigma}^{-2} \textstyle\sum_{i=1}^p Z_i^2 - 2p = \hat{\sigma}^{-2} S_p^*,$$

where

$$S_p^* = \textstyle\sum_{i=1}^p (Z_i^2 - 2\hat{\sigma}^2), \quad 1 \le p \le k;$$

and

$$\mathscr{J}_k = \min\{j\colon S_j^* = \max_{1 \le i \le k} S_i^*\}.$$

As in Section 4, the distribution of $\mathscr{J}_k$ in the special case that $\beta = 0$ provides a bound for the general case. Let

$$\Delta_{n,k}(j; \beta) = P_{\sigma,\beta}(\mathscr{J}_k > j)$$

for $0 \le j \le k$, $1 \le k < n$, $\beta \in R^k$.

THEOREM 1′. *For* $1 \le r < k$ *and* $0 \le j < k - r$,

$$\sup_{\beta \in \Omega_r - \Omega_{r-1}} \Delta_{n,k}(r + j; \beta) = \Delta_{n-r,k-r}(j; 0).$$

The proof of Theorem 1′ is similar to that of Theorem 1 and has been omitted.

Suppose now that $\beta_j = 0$ for $j = 1, \cdots, k$, so that $Z_1, \cdots, Z_n$ are i.i.d. normally distributed random variables with common mean 0 and common variance σ^2. Then, since $Z_1, \cdots, Z_k$ are independent of $\hat{\sigma}^2$, the conditional distribution of $\mathscr{J}_k$, given $\hat{\sigma}^2$, may be found from the techniques of Sections 3 and 4. Let

$$a_j(t) = P(\chi_j^2 > 2jt), \quad t > 0, j \ge 1;$$

let $p_j(t)$ and $q_j(t)$, $j \ge 0$, be defined by (4) and (5) with a_j, $j \ge 1$, replaced by $a_j(t)$, $j \ge 1$, for each $t > 0$; and define $q_\infty(t)$ by (7) with a_j replaced by $a_j(t)$ for $t > ½$. Then

$$P_{\sigma,0}(\mathscr{J}_k = j \mid \hat{\sigma}^2) = p_j(\hat{\sigma}^2/\sigma^2) q_{k-j}(\hat{\sigma}^2/\sigma^2) \tag{10}$$

for $0 \le j \le k$, $1 \le k < n$, and $\sigma^2 > 0$; and the unconditional distribution of $\mathscr{J}_k$ may be found by integrating over the possible values of $\hat{\sigma}^2$. In particular, the (unconditional) expectation of $\mathscr{J}_k$ is

$$E_{\sigma,0}(\mathscr{J}_k) = \textstyle\sum_{j=1}^k E_{\sigma,0}\{a_j(\hat{\sigma}^2/\sigma^2)\} = \sum_{j=1}^k P\{\mathscr{F}(j, n-k) > 2\},$$

where $\mathscr{F}(j, n - k)$ denotes a random variable having the F-distribution on j and $n - k$ degrees of freedom, $1 \le j \le k$.

TABLE 2
Expected Values of $\mathcal{J}_k$

k	n						
	12	24	36	48	96	192	∞
6	1.274	.873	.780	.739	.683	.657	.633
12		1.762	1.315	1.156	.972	.899	.836
18		3.712	1.952	1.526	1.132	1.003	.904
24			3.102	1.995	1.245	1.056	.926

The computations were done on an Apple II microcomputer, using formula (26.6.5) *of Abramowitz and Stegun* (1970) *to compute the F distribution function and formula* (26.4.6) *to compute the Chi squared distribution function. The last column gives the values of* $E(\mathcal{J}_k)$ *from Section* 4.

Selected values of $E_{\sigma,0}(\mathcal{J}_k)$ are listed in Table 2. Observe that the convergence of $E_{\sigma,0}(\mathcal{J}_k)$ to its limit is quite slow as $n \to \infty$.

If $n - k$ is large, then the distribution of $\hat{\sigma}^2/\sigma^2$ is nearly degenerate at 1, suggesting that the conditional probabilities in (10) might be expanded in a Taylor series about 1. This expansion is detailed in Theorem 2. First, the derivatives of $p_j(t)$ and $q_j(t)$ w.r.t. t are studied.

LEMMA 1. *For $j \geq 1$, the derivatives of $p_j(t)$ and $q_j(t)$* w.r.t. $t > 0$ *are*

$$\dot{p}_j(t) = \textstyle\sum_{i=1}^{j} \frac{1}{i}\, \dot{a}_i(t) p_{j-i}(t) \tag{11}$$

and

$$\dot{q}_j(t) = -\textstyle\sum_{i=1}^{j} \frac{1}{i}\, \dot{a}_i(t) q_{j-1}(t), \tag{12}$$

where
$$\dot{a}_i(t) = -i^{(1/2)i} t^{(1/2)i-1} e^{-it}/\Gamma(\tfrac{1}{2} i)$$

denotes the derivative of $a_i(t)$ w.r.t. $t > 0$ *for $i \geq 1$. Moreover,* (12) *holds when $j = \infty$ and $t > \frac{1}{2}$ too.*

PROOF. Let $P(t, s)$ denote the generating function of $p_j(t)$, $j \geq 0$, for $0 < s < 1$ and $t > 0$. Then $\dot{p}_j(t)$ is the coefficient of s^j in $\dot{P}(t, s) = \partial P(t, s)/\partial t$; and

$$\dot{P}(t, s) = \left\{ \textstyle\sum_{i=1}^{\infty} \frac{1}{i}\, \dot{a}_i(t) s^i \right\} P(t, s), \quad 0 < s < 1,\ t > 0.$$

Relation (11) follows immediately; and (12) may be established similarly for $1 \leq j < \infty$.

Observe that $|\dot{a}_1(t)| + |\dot{a}_2(t)| + \cdots$ is convergent uniformly in $t \geq \frac{1}{2} + \varepsilon$ for any $\varepsilon > 0$, by Stirling's Formula. Thus, for $t > \frac{1}{2}$,

$$\log q_\infty(t) = -\textstyle\sum_{j=1}^{\infty} \frac{1}{j}\, a_j(t)$$

may be differentiated term by term. That (12) holds when $j = \infty$ follows.

When iterated, (11) and (12) yield expressions for the second derivatives of $p_j(t)$ and $q_j(t)$, $j \geq 1$. For example,

$$\ddot{q}_j(t) = -\textstyle\sum_{i=1}^{j} \frac{1}{i} \{\ddot{a}_i(t) q_{j-i}(t) + \dot{a}_i(t) \dot{q}_{j-i}(t)\}$$

for $t > 0$ and $j \geq 1$; and, using the uniform convergence of $|\dot{a}_1(t)| + |\dot{a}_2(t)| + \cdots$ and $|\ddot{a}_1(t)| + |\ddot{a}_2(t)| + \cdots$, one finds easily that $\ddot{q}_j(t) \to \ddot{q}_\infty(t)$ uniformly in t on compact subintervals of $(\frac{1}{2}, \infty)$ as $j \to \infty$.

THEOREM 2. *Let $k = k_n$, $n \geq 1$, be integers for which $0 < n - k \to \infty$ as $n \to \infty$. Then*

$$P_{\sigma,0}(\mathscr{J}_k = j) = p_j q_{k-j} + \{\ddot{p}_j q_{k-j} + 2\dot{p}_j \dot{q}_{k-j} + p_j \ddot{q}_{k-j}\}\left(\frac{1}{n-k}\right) + o\left(\frac{1}{n-k}\right)$$

as $n \to \infty$ for each fixed $j \geq 0$, where $p_j = p_j(1)$, $\dot{p}_j = \dot{p}_j(1)$, etc.

PROOF. Since the probabilities are independent of σ^2, there is no loss of generality in supposing that $\sigma^2 = 1$. Let F_n denote the distribution function of $\hat{\sigma}^2$; and, for fixed $j \geq 0$, let $g_n(t) = \ddot{p}_j(t)q_{k-j}(t) + 2\dot{p}_j(t)\dot{q}_{k-j}(t) + p_j(t)\ddot{q}_{k-j}(t)$ for $t > 0$ and $n \geq 1$. Then, for any $\delta > 0$,

$$\begin{aligned} P_{\sigma,0}(\mathscr{J}_k = j) &= \int_{1-\delta}^{1+\delta} p_j(t) q_{k-j}(t) dF_n(t) + o\left(\frac{1}{n-k}\right) \\ &= p_j q_{k-j} + g_n(1)\left(\frac{1}{n-k}\right) \\ &\quad + \int_{1-\delta}^{1+\delta} \frac{1}{2}\{g_n(t_n^*) - g_n(1)\}(t-1)^2 dF_n(t) + o\left(\frac{1}{n-k}\right), \end{aligned} \tag{13}$$

where t_n^* denotes an intermediate point between t and 1. Indeed, (13) follows from elementary properties of the Chi squared distribution. Finally, it follows from Lemma 1 that g_n, $n \geq 1$, are equicontinuous in t on compact subintervals of $(\frac{1}{2}, \infty)$; so, given $\varepsilon > 0$, the last integral in (13) may be made less than $\varepsilon \int (t-1)^2 dF_n(t)$ by taking $\delta > 0$ sufficiently small; and since $\int (t-1)^2 dF(t) = 2/(n-k)$, the theorem follows.

6. Asymptotics with large n. Now consider a sequence of problems, indexed by the sample size $n \geq 1$. Suppose first that the parameter space Ω is the same for all sample sizes. Thus, Ω is an open subset of R^k for some $k \geq 1$; $0 \in \Omega$; and the nested models of interest are $\Omega_j = \{\theta \in \Omega: \theta_i = 0 \text{ for } i = j+1, \cdots, k\}$. The data $X_1, \cdots, X_n$ are assumed to have a joint density $f_n(\cdot; \theta)$ w.r.t. a dominating (sigma-finite) measure for all $\theta \in \Omega$ for each $n \geq 1$; so, the log-likelihood function may be written

$$L_n(\theta) = \log f_n(X_1, \cdots, X_n; \theta), \quad \theta \in \Omega,$$

for each $n \geq 1$. As in Section 2, the true value of the parameter is denoted by $\theta^0 = (\theta_1^0, \cdots, \theta_k^0)'$; and the index of the model selected may be written

$$J_{nk} = \min\{j: \Lambda_{nj}^* - j = \max_{0 \leq i \leq k} \Lambda_{ni}^* - i\}, \tag{14}$$

where

$$\Lambda_{ni}^* = \sup_{\theta \in \Omega_i} L_n(\theta) - L_n(\theta^0) \tag{15}$$

for $i = 1, \cdots, k$ and $n \geq 1$. It is shown that the simple normal example of Section 4 provides an asymptotic distribution for J_{nk} as $n \to \infty$, under some regularity conditions.

It is most efficient to state the regularity conditions directly in terms of the likelihood function. In their statements, P denotes a probability measure under which $X_1, \cdots, X_n$ have joint density $f_n(\cdot; \theta^0)$ w.r.t. the dominating measure. Thus, the dependence of P on n and θ^0 is supressed in the notation.

CONDITION C1. For every $\varepsilon > 0$,

$$\sup_{\|\theta - \theta^0\| \geq \varepsilon} L_n(\theta) - L_n(\theta^0) \to_p -\infty$$

as $n \to \infty$, where $\|\cdot\|$ denotes the Euclidean norm.

If a maximum likelihood estimator exists, then C1 guarantees that it converges to θ^0 in probability as $n \to \infty$.

CONDITION C2. For some $\varepsilon_0 > 0$, $L_n(\theta)$ is twice continuously differentiable in $\| \theta - \theta^0 \| < \varepsilon_0$ w.p.1 (P) for all sufficiently large n.

If C2 is satisfied, then the gradient and Hessian,

$$Z_n(\theta) = \left[\frac{\partial}{\partial \theta_1} L_n(\theta), \cdots, \frac{\partial}{\partial \theta_k} L_n(\theta) \right],$$

and

$$M_n(\theta) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_n(\theta) \colon i, j = 1, \cdots, k \right],$$

are well defined for $\| \theta - \theta^0 \| < \varepsilon_0$ and sufficiently large n. It is convenient to write $Z_n = Z_n(\theta^0)$.

CONDITION C3. Condition C2 is satisfied; and there are ε_1, $0 < \varepsilon_1 < \varepsilon_0$, positive constants α_n, $n \geq 1$, and non-random matrices M_θ, $\| \theta - \theta^0 \| \leq \varepsilon_1$ for which (i) $\alpha_n \to \infty$, as $n \to \infty$, (ii) M_θ are continuous in θ and positive definite, (iii) $Z_n/\sqrt{\alpha_n}$ is asymptotically normal with mean 0 and covariance matrix $M = M_{\theta^0}$, and (iv) $\sup_{\| \theta - \theta^0 \| \leq \varepsilon_1} \| \alpha_n^{-1} M_n(\theta) - M_\theta \|_{\mathrm{tr}} \to 0$ in probability as $n \to \infty$, where $\| \cdot \|_{\mathrm{tr}}$ denotes the trace norm.

Conditions C1, C2, and C3 imply that a maximum likelihood estimator $\hat{\theta}^n$ exists with probability approaching one, and that $\sqrt{\alpha_n}(\hat{\theta}^n - \theta^0)$ is asymptotically normal with mean vector 0 and covariance matrix M^{-1}, where $M = M_{\theta^0}$.

THEOREM 3. *Suppose that conditions* C1, C2, *and* C3 *are satisfied. Suppose also that* $\theta^0 \in \Omega_h - \Omega_{h-1}$ *for some* h, $1 \leq h < k$. *Then*

$$\lim_{n \to \infty} P(J_{nk} - h > j) = \gamma_{k-h}(j; 0)$$

for $j = 0, \cdots, k - h - 1$, *where* γ_k *is as in* (8).

Observe that $J_{nk} - h$ is the number of superfluous parameters included in the model selected.

PROOF. Since the distance from θ^0 to Ω_i is positive for all $i \leq h - 1$, it follows directly from C1 that $\max_{0 \leq i \leq h} \Lambda^*_{ni} \to_p -\infty$ as $n \to \infty$; so,

$$J_{nk} - h = \min\{ j : \Lambda^*_{n,h+j} - (h + j) = \max_{0 \leq i \leq k-h} \Lambda^*_{n,h+i} - (h + i) \}$$

with probability approaching one as $n \to \infty$. Let ε_1 be as in the statement of Condition C3; define stochastic processes $W_n(t)$, $t \in R^k$, $n \geq 1$, by

$$W_n(t) = \begin{cases} L_n\left(\theta^0 + \dfrac{t}{\sqrt{\alpha_n}}\right) - L_n(\theta^0), & \| t \| \leq \varepsilon_1 \sqrt{\alpha_n}, \\[2ex] L_n\left(\theta^0 + \dfrac{\varepsilon_1 t}{\| t \|}\right) - L_n(\theta^0), & \| t \| > \varepsilon_1 \sqrt{\alpha_n}, \end{cases}$$

and let $K_j = \{ t \in R^k : t_i = 0 \text{ for } i = j + 1, \cdots, k \}$ for $j = 1, \cdots, k$. Then for $h \leq j < k$,

$$\Lambda^*_{n,j} = \sup_{t \in K_j} W_n(t)$$

with probability approaching one as $n \to \infty$. Now, for each fixed $t \in R^k$, $W_n(t)$ converges in distribution to $W(t) = t'Z - \tfrac{1}{2} t'Mt$ as $n \to \infty$, where Z has the normal distribution with mean vector 0 and covariance matrix M; in fact, it follows from C3 that the joint distribution of $\Lambda^*_{nh} - h, \cdots, \Lambda^*_{nk} - k$ converges to that of $S_h, \cdots, S_k$ as $n \to \infty$ where

$$S_j = \sup_{t \in K_j} (t'Z - \tfrac{1}{2} t'Mt) - j$$

for $1 \leq j \leq k$. Finally, by straightforward linear algebra, there are i.i.d. standard, univariate normal random variables $Y_1, \cdots, Y_k$ for which

$$S_j = \sum_{i=1}^{j} (\tfrac{1}{2} Y_i^2 - 1), \quad j = 1, \cdots, k.$$

It follows easily that $J_{nk} - h$ converges in distribution to

$$J = \min\{j : S_{h+j} = \max_{0 \leq i \leq k-h} S_{h+i}\} = \min\{j : S_{h+j} - S_h = \max_{0 \leq i \leq k-h} S_{h+1} - S_h\},$$

which has the distribution (8), with k replaced by $k - h$.

7. Asymptotics with large n and large k. Let R^∞ denote the set of all infinite sequences of real numbers $x = (x_1, x_2, \cdots)$, endowed with the product topology; let Θ be an open subset of R^∞ for which $(0, 0, \cdots) \in \Theta$; let

$$\Omega_k = \{\theta \in \Theta : \theta_i = 0 \text{ for all } i > k\}, \quad k \geq 1;$$

and let

$$\Omega = \cup_{k=1}^{\infty} \Omega_k.$$

For each $n \geq 1$, let $X_1, \cdots, X_n$ be random vectors with joint density $f_n(\cdot; \theta)$ w.r.t. a dominating measure for some unknown $\theta \in \Omega$; let k_n, $n \geq 1$, be a non-decreasing sequence of positive integers for which $k_n \to \infty$ as $n \to \infty$; and suppose that Akaike's technique is applied with Ω replaced by Ω_{k_n} at the nth stage for each $n \geq 1$. Then the index of the model selected is

$$J_n = \min\{j : \Lambda_{nj}^* - j = \max_{0 \leq i \leq k_n} \Lambda_{ni}^* - i\},$$

where $\Lambda_{n1}^*, \Lambda_{n2}^*, \cdots$ are defined by (15). The results of Sections 4 and 6 suggest that J_n may have the limiting distribution $q_\infty p_j$, $j \geq 0$, under general conditions.

As in Section 6, the true value of the parameter is denoted by θ^0, and P denotes a probability measure under which $X_1, \cdots, X_n$ have density $f_n(\cdot; \theta^0)$ w.r.t. the dominating measure. It is assumed below that $\theta^0 \in \Omega$, and that Conditions C1, C2, and C3 are satisfied when Ω is replaced by Ω_k for all large k. In addition, the following condition is needed.

CONDITION C4. For every $\varepsilon > 0$ there is an integer $\ell_0 \geq 1$ for which

$$P\{\max_{\ell_0 \leq \ell \leq k_n} \Lambda_{n\ell}^* - \tfrac{3}{4}\ell \geq 0\} < \varepsilon \tag{16}$$

for all sufficiently large n.

THEOREM 4. *Suppose that $\theta^0 \in \Omega_h - \Omega_{h-1}$, where $h \geq 1$. Suppose also that Conditions* C1, C2, C3, *and* C4 *are satisfied. Then*

$$\lim_{n\to\infty} P(J_n - h = j) = q_\infty p_j$$

for all $j \geq 0$, *where* q_∞ *and* $p_j, j \geq 0$, *are as in* (4) *and* (5).

PROOF. Given $j \geq 0$ and $\varepsilon > 0$, there is an integer $\ell_0 \geq h$ for which $|q_{\ell_0 - j} - q_\infty| < \varepsilon$ and $P^n(J_n = J_{n\ell_0}) \geq 1 - \varepsilon$ for all sufficiently large n, where $J_{n\ell_0}$ is defined by (14). Since $\lim P^n\{J_{n\ell_0} - h = j\} = q_{\ell_0 - j} p_j$ as $n \to \infty$, by Theorem 3,

$$q_\infty p_j - 2\varepsilon \leq \liminf_{n\to\infty} P(J_n - h = j)$$
$$\leq \limsup_{n\to\infty} P(J_n - h = j) \leq q_\infty p_j + 2\varepsilon;$$

and, since $\varepsilon > 0$ was arbitrary, the theorem follows.

Condition C4 is related to the order of consistency of the maximum likelihood estimator when $k = k_n \to \infty$ with n, a difficult question. See, for example, Huber (1973) and Yohai and Maronna (1979) for discussions of this question for M estimators. The following example indicates that C4 may be replacable by a growth condition on k_n.

EXAMPLE. Suppose that n is of the form $n = km$, and that (with the obvious conventions)

$$X_{ij} = \theta_i + U_{ij}$$

where $U_{11}, \cdots, U_{km}$ are i.i.d. with a common distribution G. Suppose further that G has a positive, bounded, twice continuously differentiable density g (w.r.t. Lebesgue measure) which has finite Fisher information and satisfies some other mild conditions, described in the Appendix. Then Condition C4 is satisfied if

$$\log k = o(m) \quad \text{or} \quad k\log k = o(n) \quad \text{as } n \to \infty. \tag{17}$$

The proof of this assertion is given in the Appendix. The point of the example is this: in this simple regression problem, Condition C4 may be replaced by the mild growth condition (17).

8. Remarks. In the simple normal example of Section 3, it was shown that the probability of including no superfluous parameters is at least 0.712 and that the expected number of superfluous parameters is at most one for all values of n. This seemed reassuring—better than the author expected at the beginning of this study. These numbers seem much less reassuring in the context of Theorems 3 and 4, however, since it is possible to find the correct model with probability approaching one as $n \to \infty$. In fact, as explained below, Schwarz's (1978) Bayesian criterion will do so, under appropriate regularity conditions.

To understand the behavior of Schwarz's technique, suppose that $X_1, \cdots, X_n$ are i.i.d. random k vectors with common density

$$g_\theta(x) = \exp\{\theta' x - \psi(\theta)\}, \quad x \in R^k, \theta \in \Omega \tag{18}$$

with respect to some dominating, sigma-finite measure. Let Ω denote the natural parameter space of the family (18); suppose that Ω satisfies the conditions imposed in Section 6; and let Λ^*_{nj}, $0 \le j \le k$, be as in Section 6. Then Schwarz's technique selects the model for which $\Lambda^*_{nj} - \frac{1}{2}j \log n$ is maximum, so the index of the model selected is

$$K_n = \min(j : \Lambda^*_{nj} - \tfrac{1}{2}j \log n = \max\nolimits_{0\le i\le k} \Lambda^*_{ni} - \tfrac{1}{2}i \log n).$$

If $\theta^0 \in \Omega_r - \Omega_{r-1}$, where $r \ge 0$ and $\Omega_{-1} = \varnothing$, then Λ^*_{nj}/n converges to a negative value for $0 \le j \le r-1$ and $\Lambda^*_{nj}/n \to 0$ for $r \le j \le k$ w.p.1 as $n \to \infty$. So, $\liminf_{n\to\infty} K_n \ge r$ w.p.1. To bound the distribution of the number of superfluous parameters, observe that

$$P(K_n \ge r + i) \le \textstyle\sum_{j=r+i}^k P\{\Lambda^*_{nj} > \frac{1}{2}(j-r)\log n + \Lambda^*_{nr}\} \le \sum_{j=r+i}^k P\{\Lambda^*_{nj} > \frac{1}{2}(j-r)\log n\}$$

for $i = 1, \cdots, k-r$. For smooth exponential families, the Chi squared approximation may be applied in the tail to yield

$$P(\Lambda^*_{nj} > b \log n) \sim P(\tfrac{1}{2}\chi^2_j > b \log n) \sim \{\Gamma(\tfrac{1}{2}j)\}^{-1}(b \log n)^{1/2j-1} n^{-b} \quad \text{as } n \to \infty$$

for $i = 1, \cdots, k-r$. See Woodroofe (1978) for details. Thus,

$$P(K_n \ge r + i) = O\{n^{-1/2i}(\log n)^{1/2(r+i-1)}\} \quad \text{as } n \to \infty,$$

for $i = 1, \cdots, k-r$. The asymptotic behavior of Schwarz's technique is quite different from that of Akaike.

9. Acknowledgements. Thanks to Jan Kmenta for helpful discussions; and thanks to the editor and referees for helpful criticisms and for several of the references.

APPENDIX

The assertion made in the Example in Section 7 is proved here. The notations of Section 7 are used throughout.

THEOREM 5. *Suppose that n is of the form $n = km$, and that*

$$X_{ij} = \theta_i + U_{ij}, \quad 1 \le j \le m, 1 \le i \le k,$$

where $U_{11}, \cdots, U_{km}$ are i.i.d. *with a common distribution G. Suppose further that G has a positive, bounded, twice continuously differentiable density g* (w.r.t. *Lebesgue measure*) *for which the following conditions are satisfied*:

(i) $\mathscr{I} = \int_{-\infty}^{\infty} \frac{\{g'(x)\}^2}{g(x)} dx < \infty,$

(ii) *for some* α, $0 < \alpha < 1$, $\int_{-\infty}^{\infty} \{g(x)\}^\alpha \, dx < \infty$, *and*

(iii) *letting* $H(x) = \log g(x)$, $-\infty < x < \infty$, H'' *is bounded above,*

$\mathscr{I} = \int -H'' \, dG$, *and* $\int \sup_{|t|<\varepsilon} |H''(x-t)|^\beta \, dG(x) < \infty$ *for some* $\varepsilon > 0$ *and* $\beta > 1$. *Then Condition* C4 *holds, if* $\log k = o(m)$.

PROOF. The conditions imply that there are maximum likelihood estimators $\hat{\theta}_i = \hat{\theta}_i(X_{i1}, \cdots, X_{im})$ of θ_i for which $\hat{\theta}_1 - \theta_1, \cdots, \hat{\theta}_k - \theta_k$ are i.i.d. (with a common distribution which is independent of $\theta_1, \cdots, \theta_k$) for each fixed k and m; moreover, $\sqrt{m}\,(\hat{\theta}_1 - \theta_1)$ is asymptotically normal with mean 0 and variance $1/\mathscr{I}$ as $m \to \infty$; and, for every $\varepsilon > 0$, there is a $\rho = \rho(\varepsilon)$ for which $0 < \rho < 1$ and

$$P\{|\hat{\theta}_1 - \theta_1| \ge \varepsilon\} \le C\rho^m$$

for all $m \ge 1$ for some constant C (cf Wald, 1949).

To simplify the exposition, suppose that $\theta^0 = (0, \cdots, 0)'$ and that $\mathscr{I} = 1$. Then, for $1 \le \ell \le k$,

$$\Lambda^*_{n\ell} = \sum_{i=1}^{\ell} m\{L_i(\hat{\theta}_i) - L_i(0)\},$$

where

$$L_i(t) = \frac{1}{m} \sum_{j=1}^{m} \{H(X_{ij} - t) - H(X_{ij})\}, \quad -\infty < t < \infty.$$

Expanding $L_1, \cdots, L_k$ in Taylor series about $\hat{\theta}_1, \cdots, \hat{\theta}_k$, we find that

$$\Lambda^*_{n\ell} = -\tfrac{1}{2} \sum_{i=1}^{\ell} L_i''(t_i) Z_{mi}^2,$$

where

$$Z_{mi} = \sqrt{m}\,\hat{\theta}_i{}' \quad 1 \le i \le k,$$

and $t_1, \cdots, t_k$ are intermediate points between $\hat{\theta}_1, \cdots, \hat{\theta}_k$ and $0, \cdots, 0$. Let $\varepsilon > 0$ be so small that

$$c = \int_{-\infty}^{\infty} \sup_{|t|<\varepsilon}\{-H''(x-t)\} g(x) \, dx \le \frac{9}{8} = \frac{9}{8}\mathscr{I};$$

and let

$$A = \{|\hat{\theta}_i| \le \varepsilon, \quad \text{for all} \quad i = 1, \cdots, k\}.$$

Then

$$P(A') \le Ck\rho^m,$$

which tends to zero as $k \to \infty$ since $\log k = o(m)$. Next, the terms $L_1''(t_1), \cdots, L_k''(t_k)$ are bounded. For $i = 1, \cdots, k$, let

$$V_{mi} = \frac{1}{m} \sum_{j=1}^{m} \sup_{|t| \leq \varepsilon} - H''(X_{ij} - t), \quad W_{mi} = V_{mi} Z_{mi}^2 \cdot I_{\{|\hat{\theta}_i| \leq \varepsilon\}}.$$

Then

$$\Lambda_{n\ell}^* I_A \leq \tfrac{1}{2} \sum_{i=1}^{\ell} W_{mi}, \quad 1 \leq \ell \leq k.$$

Now $W_{m1}, \cdots, W_{mk}$ are i.i.d. for each $m \geq 1$; W_{m1} converges in distribution to $c\chi_1^2$ as $m \to \infty$, where $c \leq 9/8$, by the Central Limit Theorem and the Law of Large Numbers; and W_{m1}, $m \geq 1$, are uniformly integrable, by Lemma 2 below. In particular, the mean $\mu_m = E(W_{m1})$ converges to c as $m \to \infty$; so, there is an m_0 for which $\mu_m < 5/4$ for all $m \geq m_0$. It follows that

$$P(\max_{\ell_0 \leq \ell \leq k} \Lambda_{n\ell}^* - \tfrac{3}{4}\ell \geq 0, A) \leq P\{\max_{\ell_0 \leq \ell \leq k} \ell^{-1} \,|\, \textstyle\sum_{i=1}^{\ell} \tfrac{1}{2}(W_{mi} - \mu_m)| \geq \tfrac{1}{8}\}, \tag{19}$$

for all $m \geq m_0$. Finally, a simple adaptation of the proof of the strong Law of Large Numbers shows that the right side of (19) may be made arbitrarily small for all $m \geq m_0$ by taking ℓ_0 sufficiently large. See Lemma 3 below. Since $P(A) \to 0$ as $k, m \to \infty$, Condition C4 is satisfied.

LEMMA 2. *Suppose that* (i)–(iii) *are satisfied; then for any* $\gamma < \beta$, $\sup_{m \geq 1} E(W_{m1}^{\gamma}) < \infty$.

PROOF. For sufficiently small $\varepsilon > 0$, there are positive constants C and η and a ρ, $0 < \rho < 1$, for which

$$P(|\,\hat{\theta}_1\,| \leq \varepsilon_1 \,|\, Z_{m1}\,| > t) \leq C(e^{-\eta t^2} + \rho^m)$$

for $0 < t < \varepsilon\sqrt{m}$; see Woodroofe (1979, page 806). Thus all powers of Z_{mi} are uniformly integrable. The lemma now follows directly from (iii) and Hölder's inequality.

LEMMA 3. *For each* $m \geq 1$, *let* $Y_{m1}, \cdots, Y_{mk}$ *be* i.i.d. *random variables,* w.r.t. *probability measure* $P = P_m$, *for which*

$$E(Y_{m1}) = 0 \quad \text{and} \quad \sup_{m \geq 1} E\,|\,Y_{m1}\,|^{\alpha} < \infty \tag{20}$$

for some $\alpha > 1$. *If* $k = k_m \to \infty$ *as* $m \to \infty$, *then for every* $\delta > 0$ *there is an integer* $\ell_0 \geq 1$ *for which*

$$P(\max_{\ell_0 \leq \ell \leq k} \ell^{-1} \,|\, \textstyle\sum_{i=1}^{\ell} Y_{mi}\,| > \delta) < \delta \tag{21}$$

for all $m \geq 1$.

PROOF. Consider ℓ_0 of the form $\ell_0 = 2^q$, where $q \geq 1$ is an integer; let $r = r_m$ be an integer for which $2^{r-1} < k \leq 2^r$; and let $Y_{mi} = 0$ for $k < i \leq 2^r$. Then, the left side of (21) does not exceed

$$\textstyle\sum_{j=q}^{r} P(\max_{\ell \leq 2^j} |\,\sum_{j=1}^{\ell} Y_{mi}\,| > \tfrac{1}{2}\delta 2^j).$$

By the martingale inequality, Condition (20), and Theorem 2 of Von Bahr and Esseen (1965),

$$P(\max_{\ell \leq 2^j} |\,\textstyle\sum_{i=1}^{\ell} Y_{mi}\,| > \tfrac{1}{2}\delta 2^j) \leq (\tfrac{1}{2}\delta 2^j)^{-\alpha} E\,|\,\sum_{i=1}^{2^j} Y_{mi}\,|^{\alpha}$$

and

$$E\,|\,\textstyle\sum_{i=1}^{2^j} Y_{mi}\,|^{\alpha} \leq C \cdot 2^j$$

for all $j = 1, \cdots, r$ for some constant C. Thus, the left side of (20) does not exceed $2^\alpha \delta^{-\alpha} C \sum_{j=q}^{\infty} (\frac{1}{2})^{j(\alpha-1)}$, which is independent of m and may be made arbitrarily small by taking q sufficiently large.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1970). *Handbook of Mathematical Functions.* National Bureau of Standards.

AKAIKE, H. (1974). A new look at statistical model identification. *I.E.E.E. Trans. Auto Control* **19** 716–723.

ANDERSON, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **33** 255–265.

DANIEL, C. and WOOD, F. (1980). *Fitting Equations to Data.* Wiley, New York.

FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*, vol. 1 (3rd ed.). Wiley, New York.

FELLER, W. (1966). *An Introduction to Probability Theory and its Applications*, vol. 2. Wiley, New York.

HINKLEY, D. (1976). Note on model selection and the Akaike criterion. Unpublished manuscript, University of Minnesota.

HUBER, P. (1973). Robust regression. *Ann. Statist.* **1** 799–821.

LEHMANN, E. (1959). *Testing Statistical Hypotheses*, Wiley, New York.

MALLOWS, C. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the IMS, Manhatten, Kansas.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

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.

VON BAHR, B. and ESSEEN, G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \le r \le 2$. *Ann. Math. Statist.* **36** 299–303.

WALD, A. (1949). A note on the consistency of the maximum likelihood estimator. *Ann. Math. Statist.* **20** 596–601.

WOODROOFE, M. (1978). Large deviations of the likelihood ratio statistics with applications to sequential testing. *Ann. Statist.* **6** 72–84.

WOODROOFE, M. (1979). A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74** 799–806.

YOHAI, V. and MARONNA, R. (1979). Asymptotic behavior of M estimators for the linear model. *Ann. Statist.* **7** 258–268.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
419 SOUTH STATE STREET
ANN ARBOR, MICHIGAN 48109