On Some Criteria for Estimating the Order of a Markov Chain

Author(s): Richard W. Katz

Source: *Technometrics*, Aug., 1981, Vol. 23, No. 3 (Aug., 1981), pp. 243-249

Published by: Taylor & Francis, Ltd. on behalf of American Statistical Association and American Society for Quality

Stable URL: http://www.jstor.com/stable/1267787

TECHNOMETRICS ©, VOL. 23, NO. 3, AUGUST 1981

On Some Criteria for Estimating the Order of a Markov Chain

Richard W. Katz

Department of Atmospheric Sciences
Oregon State University
Corvallis, OR 97331

Tong (1975) has proposed a procedure for estimating the order of a Markov chain based on Akaike's information criterion (AIC). In this paper, the asymptotic distribution of the AIC estimator is derived and it is shown that the estimator is inconsistent. As an alternative to the AIC procedure, the Bayesian information criterion (BIC) proposed by Schwarz (1978) is shown to be consistent. These two procedures yield different estimated orders when applied to specific samples of meteorological observations. For parameters based on these meteorological examples, the AIC and BIC procedures are compared by means of simulation for finite samples. The results obtained have practical implications concerning whether, in the routine fitting of precipitation data, it is necessary to consider higher than first-order Markov chains.

KEY WORDS: Markov chain order; Akaike's information criterion; Bayesian information criterion; Precipitation.

1. INTRODUCTION AND SUMMARY

Several alternatives to multiple hypotheses testing techniques have recently been proposed for choosing the dimension of a model. These procedures attempt to achieve parsimony in model building, balancing the desire for a better fitting model against the desire for a model with as few parameters as possible. One such procedure, based on information theoretic concepts, is known as Akaike's information criterion (AIC) (e.g., Akaike 1974). This criterion has been applied, for example, to the problems of estimating the order of autoregressive processes (e.g., Jones 1975), autoregressive integrated moving average (ARIMA) processes (Ozaki 1977), and Markov chains (Tong 1975).

In this paper, the AIC estimator for Markov chain order is considered. Although this procedure has already been applied, for instance, to Markov chain models of meteorological data (Gates and Tong 1976; Chin 1977), little is known about the properties of the AIC procedure in this case. In Section 2, the asymptotic distribution of this estimator is derived and it is shown that the estimator is inconsistent, with a substantial probability of overestimating the true order no matter how large the sample size. This result is analogous to that obtained by Shibata (1976) for the AIC estimator of the order of an autoregressive process.

As an alternative to the AIC procedure, the Bayesian information criterion (BIC) proposed by Schwarz (1978) is extended in Section 3 to the Markov chain case and is shown to be a consistent estimator of the order of the chain. These two procedures yield different estimated orders when applied to specific samples of meteorological observations (Sec. 4). By using parameters based on these meteorological examples, the AIC and BIC procedures are compared by means of simulation for finite samples. The results obtained have practical implications concerning the routine fitting of Markov chains to precipitation data.

The model selection criteria considered in this paper are compared solely on the basis of the frequency of selection of the correct model. Such comparisons are needed, in part, because of statements in the meteorological literature (e.g., Chin 1977) indicating the belief that the AIC procedure is a consistent estimator of Markov chain order. It should be noted, however, that others (e.g., Akaike 1979; Shibata 1980) have used prediction error as a basis for comparison of model selection criteria.

2. AIC PROCEDURE

In this section, the AIC estimator for the order of a Markov chain is defined and some of its properties are examined. Only finite-state Markov chains (say s states) are considered here, and these chains are assumed to be irreducible and aperiodic. It is further assumed that the order of the Markov chain, while unknown, is known to be less than some fixed upper bound m. Formally, a kth-order Markov chain $\{X_1,$

$X_2, \ldots\}$ is characterized by transition probabilities

$$P_{i_1 \cdots i_{k+1}} = \Pr\{X_n = i_{k+1} | X_{n-1} = i_k, \ldots, X_{n-k} = i_1\},$$

where the indices $i_1, \ldots, i_{k+1}$ range over the s possible states.

Given a sample of n observations of the Markov chain $X_1, \ldots, X_n$, the AIC procedure requires the calculation of the maximized likelihood function for a kth-order chain, which is (ignoring the first few terms)

$$M_k(X_1, \ldots, X_n) = \prod_{i_1, \ldots, i_{k+1}} \hat{P}_{i_1 \cdots i_{k+1}}^{n_{i_1 \cdots i_{k+1}}}. \qquad (2.1)$$

Here, $n_{i_1 \cdots i_{k+1}}$ denotes the number of transitions from $i_1 \to i_2 \to \cdots \to i_{k+1}$ that occur in the sample, and the product in (2.1) is taken over only those indices for which $n_{i_1 \cdots i_{k+1}} > 0$. The maximum likelihood estimator of $P_{i_1 \cdots i_{k+1}}$ is given by

$$\hat{P}_{i_1 \cdots i_{k+1}} = \frac{n_{i_1 \cdots i_{k+1}}}{n_{i_1 \cdots i_k \cdot}}$$

with

$$n_{i_1 \cdots i_k \cdot} = \sum_{i_{k+1}} n_{i_1 \cdots i_{k+1}}$$

(see Billingsley 1961a for a derivation of these results).

A convenient formulation of Akaike's selection procedure in the case of Markov chain order is in terms of the likelihood ratio statistics for kth-order ($k = 0, 1, \ldots, m - 1$) versus mth-order chain $\eta_{k,m} = -2 \ln \lambda_{k,m}$ where

$$\lambda_{k,m} = \frac{M_k(X_1, \ldots, X_n)}{M_m(X_1, \ldots, X_n)}$$

(see Tong 1975).

Definition 1. The *AIC estimator* of Markov chain order, $\hat{k}_{\text{AIC}}$, is chosen such that

$$\text{AIC}(\hat{k}_{\text{AIC}}) = \min_{0 \le k < m} \text{AIC}(k)$$

where

$$\text{AIC}(k) = \eta_{k,m} - 2(s^m - s^k)(s - 1). \qquad (2.2)$$

Observe that (2.2) is just the likelihood ratio statistic modified by a penalty function (namely, twice the degrees of freedom (df)). It should also be noted that this formulation in terms of likelihood ratio statistics is equivalent to Akaike's original formulation in terms of likelihood functions. Akaike's procedure is based solely on heuristic arguments using an information theoretic concept known as the Kullback-Leibler mean information (Kullback and Leibler 1951). Tong's application of the AIC procedure to the Markov chain case is simply an extension of these same heuristic arguments. No optimality is shown and, in fact, little is known about the characteristics of this estimator, especially in the case of Markov chains.

We consider first the asymptotic distribution of $\hat{k}_{\text{AIC}}$, conditional on the true order, say k_T ($k_T = 0, 1, \ldots, m - 1$). This asymptotic distribution essentially follows from a result of Billingsley (1961b) regarding the asymptotic joint distribution of likelihood ratio test statistics for Markov chain order.

Theorem 1. Let Z_j ($j = 0, 1, \ldots, m - 1$) be independent chi-squared random variables with $s^j(s - 1)^2$ df, and set $S_k = \sum_{j=k}^{m-1} [Z_j - 2s^j(s - 1)^2]$, $k = 0, 1, \ldots, m - 1$. Then,

$$\lim_{n \to \infty} \Pr\{\hat{k}_{\text{AIC}} = k\}$$

$$= 0, \qquad \text{if } 0 \le k < k_T \qquad (2.3)$$

$$= \Pr\left\{S_k = \min_{k_T \le j < m} S_j\right\}, \qquad \text{if } k_T \le k < m. \qquad (2.4)$$

Proof. The proof of Theorem 1 is given in Appendix I.

While Theorem 1 does not give an explicit expression for the asymptotic distribution of $\hat{k}_{\text{AIC}}$, it is clear from (2.4) that the procedure is inconsistent with a positive probability of overestimating the true order. Computing the probabilities specified by (2.4) amounts to a random walk problem with nonidentically distributed random variables. Although analytical expressions have apparently not been obtained for this type of random walk problem, these probabilities can be estimated either by a recurrence relation approach requiring numerical integration or simply by direct simulation.

To illustrate the degree of inconsistency of the AIC procedure, a few examples of the evaluation of (2.4) are given for two-state (i.e., $s = 2$) Markov chains (used, for example, in meteorological applications). When $m = 3$ (a value used by Gates and Tong 1976) and $k_T = 1$, (2.4) can be expressed in closed form and the asymptotic probability of overfitting (i.e., choosing second order) is $\Pr\{\chi_2^2 > 4\} = .135$. When $m = 4$ (the value used by Chin 1977) and $k_T = 1$, (2.4) has been estimated by computer simulation (based on 10,000 replications) and the probability of overfitting (i.e., choosing second or third order) is approximately .170 (see Appendix II for details on the simulation procedures).

As should be evident from these examples, the AIC procedure has a substantial probability of overestimating the true Markov chain order. This inconsistency of the AIC procedure has previously been established in the case of estimating the order of autoregressive processes (Shibata 1976). In fact, the proof of Theorem 1 given here (see Appendix I) can easily be modified to yield Shibata's results as a special case. On the other hand, Shibata's proof apparently cannot be easily generalized to include Markov chains. The suggestion that Akaike's principle leads in general to an inconsistent estimator of the dimension of a model is also implicit in the results of Schwarz (1978).

3. BIC PROCEDURE

The question naturally arises as to whether the AIC estimator of the order of a Markov chain can be modified in some manner to produce a consistent estimator. Schwarz (1978), employing a Bayesian argument similar to that earlier used by Jeffreys (1948) (see also Stone 1979), has obtained such a modified estimator of the dimension of a model for independent and identically distributed observations whose distribution is a member of the exponential family. His procedure yields a consistent estimator of model dimension and, in addition, is asymptotically optimal (in the sense of minimizing the expected loss) under the specific assumptions made about the type of loss function and the form of prior distribution. The BIC estimator (defined later in this section) is quite similar to the AIC estimator in that it also is based on penalized likelihood functions (or, equivalently, penalized likelihood ratio statistics).

In this section the BIC procedure is applied to the problem of Markov chain order estimation. It is first shown that, unlike the AIC procedure, the BIC procedure yields a consistent estimator of Markov chain order. The asymptotic optimality of the BIC estimator is also extended to the Markov chain setting, under similar assumptions to those made by Schwarz.

Definition 2. The *BIC estimator* of Markov chain order, $\hat{k}_{\mathrm{BIC}}$, is chosen such that

$$\mathrm{BIC}(\hat{k}_{\mathrm{BIC}}) = \min_{0 \leq k < m} \mathrm{BIC}(k),$$

where

$$\mathrm{BIC}(k) = \eta_{k,m} - (s^m - s^k)(s-1)\ln \mathrm{n}. \qquad (3.1)$$

Observe that the penalty function for the likelihood ratio statistic (3.1) is the df multiplied by the natural logarithm of the sample size. It is this introduction of a function of the sample size, converging to infinity at a slow enough rate, which yields a consistent estimator of Markov chain order. In particular, comparing (2.2) and (3.1), $\hat{k}_{\mathrm{BIC}} \leq \hat{k}_{\mathrm{AIC}}$ for $n \geq 8$ (as pointed out by Schwarz 1978).

Theorem 2. $\hat{k}_{\mathrm{BIC}}$ is a consistent estimator of the true Markov chain order k_T, that is,

$$\lim_{n\to\infty} \Pr\{\hat{k}_{\mathrm{BIC}} = k_T\} = 1. \qquad (3.2)$$

Proof. The proof of Theorem 2 is given in Appendix I.

Under more restrictive conditions, the BIC procedure is not only consistent, but also asymptotically optimal (in the sense mentioned earlier). The derivation of this approximate Bayesian criterion for selecting the order of a Markov chain requires, at least formally, the specification of a prior distribution. Let $\alpha_k > 0$ ($k = 0, 1, \ldots, m-1$) be the prior probability that the chain is kth order, and let μ_k be the conditional prior distribution of the parameters (i.e., the transition probabilities) of the chain given that the chain is kth order. Here, for simplicity, μ_k is taken to be a diffuse prior for the transition probabilities. Specifically, each set of transition probabilities (holding the indices $i_1, \ldots, i_k$ fixed) is assumed to have an independent Dirichlet distribution with all parameters unity (e.g., Johnson and Kotz 1972, p. 233).

If a fixed penalty is assumed for choosing the wrong model (i.e., a $0-1$ loss function), the Bayes procedure consists of selecting the model whose posterior probability, say $BP_k(X_1, \ldots, X_n)$, is greatest. A large sample approximation for $BP_k(X_1, \ldots, X_n)$ in terms of the maximized likelihood function $M_k(X_1, \ldots, X_n)$ is now obtained.

Theorem 3. For large n,

$$\ln BP_k(X_1, \ldots, X_n) \simeq \ln M_k(X_1, \ldots, X_n) - (1/2)s^k(s-1)\ln n. \qquad (3.3)$$

Proof. The proof of Theorem 3 is given in Appendix I.

The working definition of $\hat{k}_{\mathrm{BIC}}$, namely (3.1), follows directly from Theorem 3, simply by reformulating (3.3) in terms of likelihood ratio statistics instead of likelihood functions. Although specific assumptions were made about the form of conditional prior distribution in proving Theorem 3, the approximation (3.3) should still hold under much weaker assumptions about μ_k. For example, it would be trivial to extend the proof to arbitrary Dirichlet priors. Since (3.3) does not depend on the specific prior distribution assumed, perhaps it would be more appropriate to use the terms "pseudo-Bayesian" or "quasi-Bayesian" in referring to Schwarz's procedure. It should be noted that the proof of Theorem 3 (see Appendix I) is quite similar to the arguments used by Yakowitz (1976) in deriving exact small sample Bayesian tests for Markov chain order.

4. FINITE SAMPLES

While asymptotic results have been established for both the AIC and BIC procedures (Secs. 2 and 3), it is also of importance to investigate how these estimators behave for finite samples. Analytical expressions for the exact distributions of $\hat{k}_{\mathrm{AIC}}$ and $\hat{k}_{\mathrm{BIC}}$ (as a function of the sample size n) are not available and, in any event, would probably be too complicated to be very useful. So simulation is, necessarily, relied upon to

Table 1. *Markov Chain Order Selection Applying AIC and BIC Procedures to Tel Aviv Data*

Order k	$\eta_{k,m}$	AIC(k)	BIC(k)	Approximate posterior probability
0	432.05	418.05	377.46	0.0000
1	10.11	-1.89	-36.68*	0.9945
2	4.90	-3.10*	-26.30	0.0055

*Denotes minimum.

evaluate the performance of these procedures. Attention is devoted mainly to the results for two particular examples based on meteorological data.

A two-state first-order Markov chain is considered, the possible states being labeled "0" and "1." It is convenient to characterize this process in terms of two parameters, $p = \Pr\{X_n = 1\}$ and $\rho = P_{11} - P_{01}$. Here ρ is known as the persistence parameter.

In the meteorological application, X_n represents the occurrence or nonoccurrence of precipitation at a particular location on the nth day. This model was applied by Gabriel and Neumann (1962) to Tel Aviv precipitation data. They chose a first-order chain on the basis of multiple hypotheses testing procedures. Recently, Gates and Tong (1976) reanalyzed the Tel Aviv data, applying the AIC procedure with $m = 3$ and obtaining $\hat{k}_{\text{AIC}} = 2$. When the BIC procedure is applied to the same data, again taking $m = 3$, $\hat{k}_{\text{BIC}} = 1$ is obtained (see Table 1), thus disagreeing with $\hat{k}_{\text{AIC}}$. The approximate posterior probability of a second-order model, further, is quite small (less than .01), while that of first order is virtually one.

Because of this disagreement on the appropriate order for the Tel Aviv model, it is of interest to examine the performance of $\hat{k}_{\text{AIC}}$ and $\hat{k}_{\text{BIC}}$ for simulated observations with parameters like those of the Tel Aviv data. Using the maximum likelihood estimates of the transition probabilities (assuming first-order dependence), $\hat{p} = .425$ and $\hat{\rho} = .412$ for the Tel Aviv data based on $n = 2{,}437$, observations were generated from a first-order Markov chain for these values of p and ρ (see Appendix II for details on the simulation). The maximum possible order was again taken to be two (i.e., $m = 3$).

Table 2a. *Distribution of* $\hat{k}_{\text{AIC}}$ *for First-Order Markov Chain With* $p = .425$ *and* $\rho = .412$*

Sample size	Estimated order 0	1	2
50	0.106	0.708	0.186
100	0.006	0.844	0.150
200	0.000	0.868	0.132
300	0.000	0.844	0.156
400	0.000	0.846	0.154
500	0.000	0.858	0.142
1,000	0.000	0.858	0.142
1,500	0.000	0.862	0.138
∞	0.000	0.865	0.135

*Based on 500 replications for each sample size (see Appendix II).

Table 2b. *Distribution of* $\hat{k}_{\text{BIC}}$ *for First-Order Markov Chain With* $p = .425$ *and* $\rho = .412$*

Sample size	Estimated order 0	1	2
50	0.270	0.708	0.022
100	0.038	0.950	0.012
200	0.000	0.994	0.006
300	0.000	0.996	0.004
400	0.000	0.996	0.004
500	0.000	1.000	0.000
1,000	0.000	1.000	0.000
1,500	0.000	1.000	0.000
∞	0.000	1.000	0.000

*Based on 500 replications for each sample size (see Appendix II).

Table 2a gives the estimated probability distributions of $\hat{k}_{\text{AIC}}$ for sample sizes ranging from $n = 50$ to $n = 1{,}500$ based on 500 replications. For $n \geq 100$, agreement with the asymptotic distribution (values obtained in Sec. 2) is quite close. In particular, except for $n = 50$, the AIC procedure virtually never underfits. If $m = 4$ is used instead of $m = 3$, the results (not included here) are quite similar. In this case, of course, the asymptotic probability of selecting the correct order is somewhat smaller (i.e., approximately .830 from Sec. 2).

Table 2b gives the estimated probability distributions of $\hat{k}_{\text{BIC}}$ when applied to the same generated data. For $n \geq 200$, $\hat{k}_{\text{BIC}}$ nearly always makes the correct choice of first order. For smaller sample sizes, especially $n = 50$, $\hat{k}_{\text{BIC}}$ tends to underestimate the true order much more so than does $\hat{k}_{\text{AIC}}$. Nevertheless, it does substantially better than $\hat{k}_{\text{AIC}}$ at choosing the correct order for $n \geq 100$. If $m = 4$ is used, the results (not included here) are virtually identical because $\hat{k}_{\text{BIC}}$ seldom overestimates the true order.

While the BIC procedure performs well for this example, other examples certainly exist in which it does not perform satisfactorily relative to the AIC procedure. These procedures were also applied to a set of precipitation observations for State College, Pennsylvania, previously analyzed by Katz (1977), again obtaining $\hat{k}_{\text{AIC}} = 2$ and $\hat{k}_{\text{BIC}} = 1$. Fitting a first-order Markov chain to the State College data, the parameter estimates are $\hat{p} = .380$ and $\hat{\rho} = .099$ based on $n = 1{,}120$, indicating that, although the unconditional probability of precipitation is nearly the same for both

Table 3a. *Distribution of $\hat{k}_{AIC}$ for First-Order Markov Chain With $p = .380$ and $\rho = .099$**

Sample size	Estimated order 0	1	2
50	0.716	0.164	0.120
100	0.654	0.250	0.096
200	0.494	0.416	0.090
300	0.368	0.516	0.116
400	0.290	0.586	0.124
500	0.192	0.670	0.138
1,000	0.062	0.784	0.154
1,500	0.010	0.834	0.156
∞	0.000	0.865	0.135

*Based on 500 replications for each sample size (see Appendix II).

locations, the degree of persistence is much less for State College.

Observations were generated from a first-order Markov chain for the State College estimates of p and ρ, with the simulation study being conducted in an identical manner to that for the Tel Aviv parameters. Tables 3a and 3b summarize the results. Both procedures tend to underestimate the true order of one, especially for smaller sample sizes and to a greater extent for the BIC procedure. Only for $n = 1{,}500$ does the BIC procedure select the true order more frequently than does the AIC procedure. Further, the exact distribution of $\hat{k}_{AIC}$ is only close to its asymptotic distribution for $n = 1{,}500$, whereas the exact distribution of $\hat{k}_{BIC}$ still differs somewhat from its asymptotic distribution even for $n = 1{,}500$.

A simple modification of the BIC procedure could reduce this tendency to underfit. Instead of choosing the order with maximum approximate posterior probability, the criterion would be based on the cumulative approximate posterior probability of the model being

Table 3b. *Distribution of $\hat{k}_{BIC}$ for First-Order Markov Chain With $p = .380$ and $\rho = .099$**

Sample size	Estimated order 0	1	2
50	0.920	0.066	0.014
100	0.900	0.096	0.004
200	0.820	0.178	0.002
300	0.768	0.232	0.000
400	0.714	0.286	0.000
500	0.618	0.382	0.000
1,000	0.354	0.642	0.004
1,500	0.136	0.862	0.002
∞	0.000	1.000	0.000

*Based on 500 replications for each sample size (see Appendix II).

less than or equal to a given order. The estimated order would be the lowest order whose cumulative probability is greater than some fixed threshold (e.g., .90). This criterion corresponds, at least roughly, to the Bayes procedure for a loss function that gives more weight to errors of underestimating the true order than to errors of overestimation. Moreover, it is easy to see that this modified version of the BIC procedure is asymptotically equivalent to the original criterion—in particular, retaining the same asymptotic properties such as consistency.

The performance of this modified BIC estimator of Markov chain order has been evaluated in a similar manner, applying it to the same simulated data (detailed results not included here). In this evaluation the approximate posterior probability of underfitting was required, as an ad hoc choice, to be less than .10 (in other words, a cumulative probability threshold of .90). For the first-order chain with the State College parameters, the modified BIC procedure is superior to the original criterion in terms of decreasing the probability of underfitting and to the AIC procedure in terms of increasing the probability of choosing the true order for small sample sizes. At the same time, it still performs virtually as well as the original procedure for the first-order Markov chain with the Tel Aviv parameters.

5. CONCLUDING REMARKS

Gates and Tong (1976) have claimed, on the basis of the AIC procedure, that the Tel Aviv precipitation data should be fitted using a second-order Markov chain. We, however, have established that first order is adequate. Chin (1977) fitted Markov chains to precipitation data at more than 100 different locations in the United States and, again using the AIC technique, found that higher than first-order chains were sometimes necessary. We suspect that higher than first-order chains need not be considered, based on our results.

As mentioned in Section 1, prediction error could also be used as a basis for comparison of model selection criteria. Because simulation studies are necessarily based on the assumption that the data are generated by a member of the class of models under consideration (i.e., a finite-order Markov chain), they cannot be expected to identify which criterion will, when applied in practice to meteorological observations, select models that result in the smallest prediction error. Genuine prediction tests, using actual meteorological observations, will be necessary to resolve this issue.

6. ACKNOWLEDGMENTS

A portion of this research was conducted while the author was a scientist in the Environmental and Societal Impacts Group at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation. I am grateful for the comments of the referees, an associate editor, and the editor.

7. APPENDIX I

Proof of Theorem 1. (a) (2.3) holds since the $\eta_{k,m}$ are asymptotically degenerate at ∞ (i.e., the power of the test converges to one) for $k < k_T$ (e.g., Anderson and Goodman 1957). (b) To verify that (2.4) holds, first note that by the definition of the likelihood ratio statistic

$$\eta_{k,m} = \eta_{k,k+1} + \cdots + \eta_{m-1,m}. \tag{7.1}$$

Further, Theorem 6.3 of Billingsley (1961b) establishes that

$$(\eta_{k_T,k_T+1}, \ldots, \eta_{m-1,m}) \to (Z_{k_T}, \ldots, Z_{m-1}) \quad \text{in distribution,} \tag{7.2}$$

with the Z_j's as defined in the statement of the theorem.

By (7.1) and Definition 1, $\hat{k}_{\text{AIC}}$ can be expressed as a function, say h, of $\eta_{k_T,k_T+1}, \ldots, \eta_{m-1,m}$ (because (2.3) holds, $\eta_{k,k+1}$, $k < k_T$, can be ignored). Specifically,

$$h(z_1, \ldots, z_q) = i + k_T - 1 \qquad \text{if } z_i^* = \min_{1 \le j \le q} z_j^*,$$

where

$$z_i^* = \sum_{j=i}^{q} [z_j - 2s^{j+k_T-1}(s-1)^2]$$

and

$$q = m - k_T.$$

Now, by Theorem 5.1 of Billingsley (1968), if h were continuous almost everywhere with respect to the probability measure associated with $(Z_{k_T}, \ldots, Z_{m-1})$, then (7.2) implies

$$\hat{k}_{\text{AIC}} \to h(Z_{k_T}, \ldots, Z_{m-1}) \text{ in distribution,}$$

which is (2.4). That h is continuous almost everywhere follows, since h is discontinuous only when one or more equalities occur among the z_i's and these ties have asymptotically probability zero of occurring.

Proof of Theorem 2. A needed result concerning the rate of the convergence of likelihood ratio statistics for tests of Markov chain order (Anderson and Goodman 1957) is the following:

$$(1/n)\eta_{k,m} \to c_k \ne 0 \text{ in probability,} \qquad \text{if } 0 \le k < k_T. \tag{7.3}$$

Now (3.3) and the fact that the penalty function in (3.1), namely $(s^m - s^k)(s-1)\ln n$, is a decreasing function of k yield (3.2).

Proof of Theorem 3. If we ignore the first few terms of the likelihood function (as in the formulation of the AIC procedure), Bayes' formula and the properties of the Dirichlet distribution (see Johnson and Kotz 1972) give

$$BP_k(X_1, \ldots, X_n) \propto \int \alpha_k \prod_{i_1, \ldots, i_{k+1}} P_{i_1 \cdots i_{k+1}}^{n_{i_1 \cdots i_{k+1}}} \, d\mu_k$$
$$\propto \prod_{i_1, \ldots, i_k} \frac{\prod_{i_{k+1}} (n_{i_1 \cdots i_{k+1}})!}{(n_{i_1 \cdots i_k \cdot} + s - 1)!}. \tag{7.4}$$

Using (2.1) and (7.4),

$$[BP_k(X_1, \ldots, X_n)]/[M_k(X_1, \ldots, X_n)]$$
$$\propto \prod_{i_1, \ldots, i_k} \frac{n_{i_1 \cdots i_k \cdot}^{\,n_{i_1 \cdots i_k \cdot}}}{(n_{i_1 \cdots i_k \cdot} + s - 1)!} \times \prod_{i_{k+1}} \frac{(n_{i_1 \cdots i_{k+1}})!}{n_{i_1 \cdots i_{k+1}}^{\,n_{i_1 \cdots i_{k+1}}}}. \tag{7.5}$$

Substituting Stirling's formula in (7.5) and taking logarithms yields (3.3).

8. APPENDIX II

The simulations and computations reported in Sections 2 and 4 were carried out on a CDC 7600 computer. Uniform pseudorandom numbers were generated using the multiplicative congruential method (CDC function RANF). As a precautionary measure, before using these generated numbers in the simulations, they were also shuffled (one of the modifications recommended by Chambers 1977, p. 173).

The simulation results reported in Section 2 required the generation of chi-squared pseudorandom numbers. These were generated by employing the representation of a chi-squared random variable as a sum of exponentials (and one squared normal if the df are odd). From the uniform pseudorandom numbers, the exponentials were obtained by applying the logarithmic transformation and the normals by applying the so-called Box-Muller transformation.

[*Received May 1980. Revised January 1981.*]

REFERENCES

AKAIKE, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

——— (1979), "A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting," *Biometrika*, 66, 237–242.

ANDERSON, T. W., and GOODMAN, L. A. (1957), "Statistical Inference About Markov Chains," *Ann. Math. Statist.*, 28, 89–110.

BILLINGSLEY, P. (1961a), "Statistical Methods in Markov Chains," *Ann. Math. Statist.*, 32, 12–40.

——— (1961b), *Statistical Inference for Markov Processes*, Chicago: University of Chicago Press.

BILLINGSLEY, P. (1968), *Convergence of Probability Measures*, New York: John Wiley.

CHAMBERS, J. M. (1977), *Computational Methods for Data Analysis*, New York: John Wiley.

CHIN, E. H. (1977), "Modeling Daily Precipitation Occurrence Process With Markov Chain," *Water Resources Research*, 13, 949–956.

GABRIEL, K., and NEUMANN, J. (1962), "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv," *Quarterly Journal of the Royal Meteorological Society*, 88, 90–95.

GATES, P., and TONG, H. (1976), "On Markov Chain Modeling to Some Weather Data," *Journal of Applied Meteorology*, 15, 1145–1151.

JEFFREYS, H. (1948), *Theory of Probability* (2nd ed.), Oxford: Clarendon Press.

JOHNSON, N. L., and KOTZ, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, New York: John Wiley.

JONES, R. H. (1975), "Fitting Autoregressions," *J. Amer. Statist. Assoc.*, 70, 590–592.

KATZ, R. W. (1977), "Precipitation as a Chain-Dependent Process," *Journal of Applied Meteorology*, 16, 671-676.

KULLBACK, S., and LEIBLER, R. A. (1951), "On Information and Sufficiency," *Ann. Math. Statist.* 22, 79–86.

OZAKI, T. (1977), "On the Order Determination of ARIMA Models," *Applied Statistics*, 26, 290–301.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

SHIBATA, R. (1976), "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117–126.

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

STONE, M. (1979), "Comments on Model Selection Criteria of Akaike and Schwarz," *J. Roy. Statist. Soc.*, Ser. B, 41, 276–278.

TONG, H. (1975), "Determination of the Order of a Markov Chain by Akaike's Information Criterion," *Journal of Applied Probability*, 12, 488–497.

YAKOWITZ, S. J. (1976), "Small Sample Hypothesis Tests of Markov Order, With Applications to Simulated and Hydrologic Chains," *J. Amer. Statist. Assoc.*, 71, 132–136.