J. R. Statist. Soc. B (1985)
47, *No.* 3, *pp.* 528–539

A Model for High-order Markov Chains

By ADRIAN E. RAFTERY†

Trinity College, Dublin, Ireland

[Received August 1983. Revised July 1984]

SUMMARY

A model for Markov chains of order higher than one is introduced which involves only one additional parameter for each extra lag. Asymptotic properties and the autocorrelation structure are investigated. Three examples are given in which the model appears to model data more successfully than both the usual high-order Markov chain and the alternative models of Jacobs and Lewis (1978), Pegram (1980) and Logan (1981).

Keywords: DISCRETE-VALUED TIME-SERIES; LIMIT THEOREM; YULE-WALKER EQUATIONS; BIC; WIND POWER; INTERPERSONAL RELATIONSHIPS; OCCUPATIONAL MOBILITY

1. INTRODUCTION

The purpose of this paper is to introduce a model for lth-order Markov chains which combines realism with parsimony. The conventional model for lth-order Markov chains has $(m-1)m^l$ parameters where m is the number of states. This large number of parameters has discouraged use of the model for $l > 1$, even when higher-order dependence *is* present.

Our model is as follows. Consider a sequence $\{X_t: t \in N\}$, where N is the set of positive integers, taking values in $\{1, 2, \ldots, m\}$. The model is

$$P\,[X_t = j_0 \mid X_{t-1} = j_1, \ldots, X_{t-l} = j_l] = \sum_{i=1}^{l} \lambda_i q_{j_0 j_i}, \tag{1.1}$$

where $\lambda_1 + \ldots + \lambda_l = 1$ and $Q = \{q_{jk}\}$ is a non-negative $m \times m$ matrix with column sums equal to 1, such that

$$0 \leqslant \sum_{i=1}^{l} \lambda_i q_{jk_i} \leqslant 1 \; (j, k_1, \ldots, k_l = 1, \ldots, m). \tag{1.2}$$

Thus the conditional probability of observing $X_t = j_0$ given the past is a linear combination of contributions from each of $X_{t-1}, \ldots, X_{t-l}$. In addition, (1.1) is analogous to the standard $AR(l)$ model in that:

(i) each additional lag, after the first, is specified by a single parameter; and
(ii) the autocorrelations satisfy a system of linear equations similar to the Yule-Walker equations; see Section 2.2.

Another way of writing (1.1) is as follows. Let $\chi_t = (x_t(1), \ldots, x_t(m))'$, where $x_t(j) = 1$ if $X_t = j$ and 0 otherwise. Let $\hat{\chi}_t = (\hat{x}_t(1), \ldots, \hat{x}_t(m))'$ where the random variable $\hat{x}_t(j)$ is a function of past values and is realised as the conditional probability $P\,[X_t = j \mid X_{t-1} = j_1, X_{t-2} = j_2, \ldots]$

† *Present address*: Dept of Statistics GN-22, University of Washington, B313 Padelford Hall, Seattle, WA 98195, USA.

0035–9246/85/47528 $2.00

when $X_{t-1} = j_1, X_{t-2} = j_2, \ldots$. Then (1.1) can be written

$$\hat{\chi}_t = \sum_{i=1}^{l} \lambda_i Q \chi_{t-i}. \tag{1.3}$$

When $l = 1$, (1.3) defines a first-order Markov chain with transition matrix Q (using Bartlett's, 1978, convention whereby the columns, and not the rows, of a transition matrix sum to 1). If $Q\pi = \pi$ where $\pi = (\pi_1, \ldots, \pi_m)'$ is a positive m-vector with $\pi_1 + \ldots + \pi_m = 1$, then the equilibrium distribution of the process is π; see Section 2.1.

We know of only three other models for high-order Markov chains. One was proposed by Pegram (1980) and Jacobs and Lewis (1978c), who called it the *DAR*(l) process, and generalizes a model of Lloyd (1977) and Jacobs and Lewis (1978a, b). It is a special case of (1.3) with

$$Q = \theta I + (1 - \theta)\pi 1', \tag{1.4}$$

where I is the identity matrix of order m, and 1 is an m-vector of ones. (1.4) does achieve parsimony, but the range of dependence patterns that can be represented is severely restricted when $m \geqslant 3$. This is because X_t depends on X_{t-1}, say, only through the probability of their being *equal*; the model does not allow X_t to have a high probability of taking values "close to" X_{t-1}. For instance, suppose X_t is the contents of a Lloyd reservoir on day t, taking values in $\{0, 1, 2, 3\}$. Then the conditional probability that $X_t = 0$, given the past, is the same if $X_{t-1}, \ldots, X_{t-l}$ were all 3 as if they were all 1. Thus the reservoir would be as likely to be empty on day t if it were full on the preceeding days as if it were nearly empty, which may be unrealistic. Non-Markovian models similar in conception to (1.4) have been proposed by Jacobs and Lewis (1978a, b, 1983). For these, in addition to the other problems, efficient estimation is difficult because the likelihood cannot be easily written down.

When $m \geqslant 3$, (1.1) can not only represent a much wider range of dependence patterns than (1.4), but it can also capture a much wider range of autocorrelations without any additional parameters; see Section 2.3. When $m = 2$, however, (1.1) is not a generalization of (1.4). Thus this paper contributes nothing new in the binary case, which has in any event already been studied in detail. Models similar to (1.4) have been proposed by Kanter (1973), Klotz (1975) and McKenzie (1981), while Kedem (1980) describes an alternative approach.

The other two models are due to Logan (1981), who describes them as constrained and unconstrained, respectively. They involve rather a large number of parameters. In Table 1 we give the numbers of parameters for the models mentioned here for common values of l and m.

TABLE 1
Numbers of parameters for Markov chain models

		Model				
l	*m*	(1.4)	(1.1)	*LC*	*LU*	*Usual*
2	2	3	3	4	4	4
	3	4	7	12	15	18
	4	5	13	20	28	48
	5	6	21	30	45	100
3	2	4	4	8	8	8
	3	5	8	18	33	48
	4	6	14	28	76	192
	5	7	22	40	145	500

(*LC*: Logan's constrained model. *LU*: Logan's unconstrained model.)

In Section 3 we give three examples in which (1.1) appears to combine parsimony and realism more satisfactorily than the alternatives.

2. PROPERTIES

The results of this section are proved in the Appendix.

2.1. *The Limit Theorem*

We now state the basic limit theorem for (1.1).

Theorem 1. Suppose that $\{X_t : t \in N\}$ is defined by (1.1), that Q is positive, and that $\pi = (\pi_1, \ldots, \pi_m)'$ is such that $\pi_i > 0$ $(i = 1, \ldots, m)$, $\pi_1 + \ldots + \pi_m = 1$ and $Q\pi = \pi$. Then

$$\lim_{t \to \infty} P[X_t = j \mid X_l = i_l, \ldots, X_1 = i_1] = \pi_j \qquad (i_1, \ldots, i_l, j = 1, \ldots, m).$$

It is apparent from the proof of Theorem 1 that if $Z_t = (X_t, \ldots, X_{t+l-1})'$ then $\{Z_t\}$ is an ergodic Markov chain with state space $\{1, \ldots, m\}^l$ and equilibrium distribution ξ. Thus if Z_1 has the distribution ξ, then $\{Z_t\}$ is stationary and so is $\{X_t\}$. In what follows we assume this to be the case.

2.2. *Bivariate Distributions and Autocorrelations*

The autocorrelation structure of (1.1) does not satisfy the Yule-Walker equations in general, which is not surprising since the same is true of the usual first-order Markov chain, of which (1.1) is a generalization. However, as the following theorem shows, the entire bivariate distribution does satisfy a system of linear equations similar to the Yule-Walker equations.

Theorem 2. Suppose $\{X_t : t \in N\}$ is defined by (1.1) and is stationary. Let $P(k)$ be an $m \times m$ matrix with elements

$$p_{ij}(k) = P[X_{t+k} = i, X_t = j] \quad (i, j = 1, \ldots, m; k \in Z),$$

and $P(0) = \operatorname{diag}\{\pi_1, \ldots, \pi_m\}$. Then

$$P(k) = \sum_{g=1}^{l} \lambda_g Q P(k-g) \quad (k \in N). \tag{2.1}$$

Note that when $l = 1$, (2.1) reduces to $P(k) = Q^k P(0)$. This is the standard result for first-order Markov chains. For (1.4), (2.1) becomes

$$p_{ih}(k) = \theta \sum_{j=1}^{l} \lambda_j p_{ih}(k-j) + (1-\theta)\, \pi_i \pi_h$$

so that each sequence $\{p_{ih}(k) : k \in N\}$ does satisfy a system of Yule-Walker equations in that case.

From Theorem 2 we can derive a system of equations for the autocorrelations themselves that resembles the Yule-Walker system, although it does not in general allow them to be calculated uniquely (sometimes it does: see Section 2.3).

Corollary to Theorem 2. Suppose that $\{X_t : t \in N\}$ is defined by (1.1) and is stationary. Suppose Y_t is a random variable with the distribution $Q\chi_t$, i.e. the conditional distribution of Y_t given $\chi_t = x$ is Qx, so that $P[Y_t = i \mid X_t = j] = q_{ij}$. Let $\rho_k = \operatorname{corr}(X_{t+k}, X_t)$ and $\tilde{\rho}_k = \operatorname{corr}(Y_{t+k}, X_t)$. Then

$$\rho_k = \sum_{j=1}^{l} \lambda_j \tilde{\rho}_{k-j} \quad (k \in N). \tag{2.2}$$

We now investigate the uniqueness of the solution to (2.1). We note that $P(-k) = P(k)'$ and also that if $P(1), \ldots, P(l-1)$ are known then $P(l), P(l+1), \ldots$ are determined uniquely. The question is thus whether the $(l-1)m^2$ linear equations (2.1) with $k = 1, \ldots, l-1$ in the $(l-1)m^2$ unknowns $p_{ih}(k)$ $(i, h = 1, \ldots, m; k = 1, \ldots, l-1)$ have a unique solution. In the following theorem we give sufficient conditions for this to be the case.

Theorem 3.

(i) If $l = 2$, (2.1) has a unique solution if $0 < \lambda_1 \leqslant 1$.
(ii) If $l = 3$, (2.1) has a unique solution if $\lambda_i \geqslant 0$ $(i = 1, 2, 3)$ and Q has at least one row all of whose elements are non-zero.
(iii) If $l \geqslant 4$, (2.1) has a unique solution if $\lambda_i \geqslant 0$ $(i = 1, \ldots, l)$ and

$$(1 - \eta_1 - \ldots - \eta_m)(2 - \lambda_1 - \lambda_{l-1} - \lambda_l) < 1, \tag{2.3}$$

where $\eta_i = \min\{q_{i1}, \ldots, q_{im}\}$.

2.3. *The Autocorrelation Structure with Three States*

We now examine in greater detail the correlation structure when $m = 3$. For simplicity we consider only the special case where the marginal probabilities are equal, i.e. $\pi = \frac{1}{3}1$, and where Q is chosen in such a way that the autocorrelations do satisfy a set of Yule–Walker equations. We let

$$Q = \tfrac{1}{3}(1 - |\alpha|)J + \begin{cases} \alpha I, & 0 \leqslant \alpha \leqslant 1 \\ |\alpha| E, & -1 \leqslant \alpha < 0 \end{cases} \tag{2.4}$$

where $|\alpha| \leqslant 1$, J is a 3×3 matrix of ones, I is the identity matrix of order 3 and

$$E = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

When $\alpha \geqslant 0$ this is exactly (1.4); the inclusion of the $\alpha < 0$ case, which is possible only in the context of (1.1), broadens (2.4) considerably without introducing any extra parameters.

We now investigate the range of autocorrelations that can be represented by (2.4) when $l = 2$. The equation (2.2) implies that

$$\begin{cases} \rho_1 = \phi_1 + \phi_2 \rho_1 \\ \rho_2 = \phi_1 \rho_1 + \phi_2 \end{cases}, \tag{2.5}$$

where $\phi_1 = \lambda_1 \alpha$, $\phi_2 = \lambda_2 \alpha$.

We first derive the range of possible values of (ρ_1, ρ_2) when $\alpha \geqslant 0$, i.e. for (1.4).

Combining (2.5) with the constraints (1.2) yields the admissible region

$$\begin{aligned} &\rho_1 \geqslant -\tfrac{1}{2} \\ &\rho_1 + \rho_2 \geqslant 0 \\ &\rho_2 \geqslant \{\rho_1(1 + 3\rho_1) - 1\}/(2 + \rho_1). \end{aligned} \tag{2.6}$$

This region is illustrated in Fig. 1(a). We see that the range of possible autocorrelations is rather limited. In particular $\rho_1 \geqslant -\frac{1}{2}$ and $\rho_2 \geqslant -\frac{1}{4}$ which is disturbing given that the standard first-order Markov chain allows ρ_1 to take any negative value greater than -1 in this case.

In order to see to what extent the range of autocorrelations represented by (1.1) is more complete than for (1.4) we carry out the same calculations for $\alpha < 0$. This yields

$$\begin{cases} -(1 + 2\rho_1) \leqslant \rho_2 \leqslant -\rho_1 \\ \rho_2(1 + 2\rho_1) \geqslant 2\rho_1(1 + \rho_1) - 1 \\ \rho_2 - 1 \leqslant (\rho_1 + 1)(\rho_1 - \rho_2). \end{cases} \tag{2.7}$$

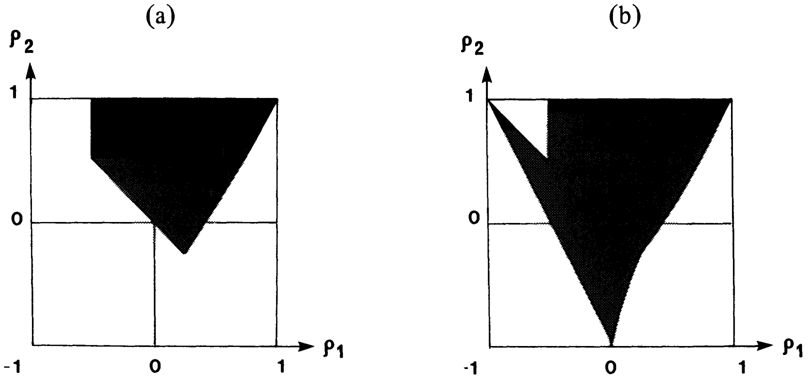

Fig. 1. Range of possible autocorrelations for the three-state model specified by (a) equation (14); (b) equation (1.1).

The admissible region for the full model, i.e. the union of the regions (2.6) and (2.7), is shown in Fig. 1(b). We see that using (1.1) instead of (1.4) extends the set of possible autocorrelations considerably, and that both ρ_1 and ρ_2 can take any value between -1 and 1. A comparison of Fig. 1(b) with Fig. 3.3(b) of Box and Jenkins (1976) shows that the range of autocorrelations for (1.1) is almost as great as for the standard $AR(2)$ model, in spite of the severity of the restrictions placed on the parameters of (1.1).

3. EXAMPLES

3.1. *Estimation and Model Choice*

We first describe the statistical procedures used. Maximum likelihood estimates of the parameters of (1.1) were obtained by numerically maximizing the log-likelihood

$$L = \sum_{i_0, i_1, \ldots, i_l = 1}^{m} n_{i_0}, \ldots, {}_{i_l} \log \left(\sum_{j=1}^{l} \lambda_j q_{i_0 i_j} \right),$$

where $n_{i_0}, \ldots, {}_{i_l} = \Sigma_t x_t(i_0)\, x_{t-1}(i_1) \ldots x_{t-l}(i_l)$, subject to (1.2). A constrained non-linear optimization program was used.

To compare models we used an information criterion rather than a multiple hypothesis testing procedure, because the models are not nested. Tong (1975) recommends choosing the model which minimizes $AIC = -2L + 2k$, where k is the number of independent parameters. However, Katz (1981) preferred the alternative of choosing the model which minimizes $BIC = -2L + k \log n$ because (i) it is a consistent estimator of Markov chain order, unlike the AIC method; (ii) it is approximately the same as choosing the model with highest posterior probability; (iii) it chooses simpler models; and (iv) it performed well in a simulation experiment. We therefore report our results in terms of BIC. However, the model "choices" made are the same as those resulting from an appropriate sequence of likelihood-ratio tests at the 5 per cent level.

3.2. *Wind Power*

In order to investigate wind turbine design, hourly windspeeds were classified into one of four states defined by the mode of operation of a particular turbine. State 1 corresponds to no power being produced (0 to 8 knots), state 2 to the turbine outputting less power than its full potential (8 to 16 knots), state 3 to its operating at full capacity (16 to 25 knots), and state 4 to its being closed down due to excessively high winds (over 25 knots). We consider a sequence of 672 hourly windspeeds at Belmullet, Ireland from 1st to 28th July, 1962. For a full description of this and related data see Raftery *et al.* (1982).

The "time-of-day" effect is negligible and there is no seasonal effect over this short period, so that a stationary model is appropriate. The usual Markov chain model, as well as (1.1) and (1.4), were fitted to the data for orders 0, 1, 2, 3, 4. All the model comparisons were carried out "within" the fourth-order model, so that the contributions to L from the first four observations were ignored in all cases. Jumps of more than one state were not observed and parameters corresponding to them were taken to be identically zero.

TABLE 2

BIC values for wind power data

	Markov chain		*(1.1)*		*(1.4)*	
Order	*k*	*BIC*	*k*	*BIC*	*k*	*BIC*
0	3	1758.55	-	-	-	-
1	6	874.09	-	-	-	-
2	16	874.47	7	836.73	5	915.34
3	42	1006.06	8	828.05*	6	892.41
4	110	1410.45	9	834.56	7	896.73

k = number of parameters, * indicates lowest *BIC* value

From Table 2, (1.1) with $l = 3$ is the model with smallest *BIC*, the only near competitors being (1.1) with $l = 2$ and 4. The same model is chosen if one uses a sequence of likelihood ratio tests at the 5 per cent level, and if one uses *AIC*. The parameter estimates are $\hat{\lambda}_1 = 0.629$, $\hat{\lambda}_2 = 0.206$, $\hat{\lambda}_3 = 0.165$ and

$$\hat{Q} = \begin{pmatrix} 0.837 & 0.058 & 0 & 0 \\ 0.163 & 0.854 & 0.113 & 0 \\ 0 & 0.088 & 0.847 & 0.116 \\ 0 & 0 & 0.040 & 0.884 \end{pmatrix}$$

(1.4) is clearly inadequate and even the usual first-order Markov chain seems better.

3.3. *Inter-personal Relationships*

Katz and Proctor (1959, Table 2) give 300 two-step transitions for relationships between students at two-month intervals, the three states being "mutual", "one-way", and "indifferent". The usual likelihood-ratio test for Markov chain order rejects the first-order hypothesis, and so we compare only the second-order models. These data were also analysed by Bishop *et al.* (1975, chapter 7) and in accordance with their widely-used conventions for counting degrees of freedom we do not count parameters which correspond to estimated zeros in the one-step transition matrix.

TABLE 3

BIC values for Katz and Proctor's data (* indicates lowest *BIC* value)

Model	*k*	*BIC*
Second-order Markov chain	12	326.63
(1.1) with $l = 2$	5	300.02*
(1.3) with $l = 2$	4	316.76

From Table 3, (1.1) has the smallest *BIC* by a considerable margin over (1.4). The parameter estimates are $\hat{\lambda}_1 = 0.754$, $\hat{\lambda}_2 = 0.246$ and

$$\hat{Q} = \begin{pmatrix} 0.581 & 0.133 & 0 \\ 0 & 0.545 & 0.093 \\ 0.419 & 0.322 & 0.907 \end{pmatrix}$$

3.4. *Occupational Mobility of Physicists*

We reanalyse the data used by Logan (1981) to illustrate his own models. His Table 1 gives 9170 two-step transitions between the three states "Management", "Research", and "Teaching" for Ph.D. physicists in the U.S. As in Section 3.3 the usual likelihood-ratio test rejects the first-order hypothesis, so again we compare only the second-order models. The *BIC* values are given in Table 4, those for Logan's models being based on his own tables of expected values.

TABLE 4
BIC values for Logan's data (*indicates lowest *BIC* value)

Model	*k*	*BIC*
Second-order Markov chain	18	10702.69
(1.1) with $l = 2$	7	10618.11*
(1.4) with $l = 2$	4	10645.18
Logan constrained (second-order)	12	10652.52
Logan unconstrained (second-order)	15	10676.64

In Table 4, (1.1) has the smallest *BIC*, by a substantial margin. The parameter estimates are $\hat{\lambda}_1 = 0.711, \hat{\lambda}_2 = 0.289$ and

$$\hat{Q} = \begin{pmatrix} 0.900 & 0.104 & 0.071 \\ 0.069 & 0.830 & 0.078 \\ 0.031 & 0.065 & 0.851 \end{pmatrix}.$$

Logan's models both fare rather poorly, with higher *BIC* values than both (1.1) and (1.4), in spite of their having been derived specifically for this kind of data. This seems to be because they involve more parameters than are justified by the fit obtained.

4. CONCLUSION

We have introduced a model for Markov chains of order higher than one which involves only one additional parameter for each extra lag, can represent a wide range of dependence structures, and appears to model real data successfully.

It is limited in being defined only for a finite state space. However, discrete-valued time series often have an infinite state space, even if almost all the observations are small in practice. An important example of this is when the observations are counts of events in a point process. In this case consecutive counts will not be independent unless the events form a Poisson process. Unless a scientifically relevant model is available, the difficulty of fitting models for stationary point processes when only counts are available may make it worthwhile modelling the sequence of counts directly.

The main problem in generalizing (1.3) to an infinite state space is that the matrix Q becomes infinite. Thus for the model to have an operational meaning Q must be specified by a finite number of parameters. A simple way of deriving such a model for Q is to consider a random vector (Y, Z) with the desired marginal distribution and to define $q_{jk} = P\,[Y = j \mid Z = k]$. Thus, for example, we may obtain a Poisson model by taking (Y, Z) to have the bivariate Poisson distribution of Holgate (1964), and a negative binomial model by taking (Y, Z) to have the bivariate negative binomial distribution described by Johnson and Kotz (1969, Section 11.3).

One may build models with a richer structure than (1.1). For example, the "autoregressive" form of (1.3) suggests adding in "moving average" terms to construct a class of models with a form analogous to that of the standard *ARMA* class, which would generalize the *DARMA* models of

Jacobs and Lewis (1978a, b, 1983). However, such models would not be Markovian in general, and efficient parameter estimation could be a major problem.

Another possibility is to allow Q to vary with lag in (1.3). While this would in general be overparameterized for most purposes, it could be a valuable generalization if Q itself is modelled.

The success of log-linear models for discrete multivariate data might lead us to consider applying them to discrete-valued sequences, as an alternative to the linear models with which this paper is concerned. This could be done by subjecting the parameters of the log-linear formulation of the lth-order Markov chain, given in Bishop *et al.* (1975, chapter 7), to constraints. Such an approach may have certain disadvantages compared with the linear models discussed in this paper. For example, the marginal distribution is a complicated function of all the parameters. Also, the generalization to an infinite state space, which, as we saw, is simple for linear models, seems less straightforward.

In seeking a parsimonious model, we have largely confined our attention to simplifying the effect of increasing lags. The number of parameters could be further reduced by modelling the matrix Q. Various ways of doing this are suggested in this paper. In Section 2.3 we saw how, in a special case, considering only second-order properties leads naturally to a model for Q, while in Section 3.2 the data themselves impose restrictions on Q. We have also considered how modelling Q enables us to deal with an infinite state space. When the state space is finite, the problem of modelling Q is equivalent to that of modelling square contingency tables with marginal homogeneity, for which the work of Clogg (1982) and others could be useful.

ACKNOWLEDGEMENTS

I would like to thank the Editor, a referee, Dermot Duggan and Kris Mosurski for very helpful comments. Part of this research was done during a visit to the Technische Universität München financed by the Deutscher Akademischer Austauschdienst.

REFERENCES

Bartlett, M. S. (1978) *An Introduction to Stochastic Processes*, 3rd ed. Cambridge: University Press.

Bellman, R. (1970) *Introduction to Matrix Analysis*, 2nd ed. New York: McGraw-Hill.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press.

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

Clogg, C. C. (1982) Some models for the analysis of association in multiway cross-classifications having ordered categories. *J. Amer. Statist. Ass.*, 77, 803–815.

Cox, D. R. and Miller, H. D. (1965) *The Theory of Stochastic Processes*. London: Methuen.

Holgate, P. (1964) Estimation for the bivariate Poisson distribution. *Biometrika*; **51**, 241–245.

Jacobs, P. A. and Lewis, P. A. W. (1978a) Discrete time series generated by mixtures I: Correlational and runs properties. *J. R. Statist. Soc.* B, **40**, 94–105.

——— (1978a) Discrete time series generated by mixtures II: Asymptotic properties. *J. R. Statist. Soc.* B, **40**, 222–228.

——— (1978c) Discrete time series generated by mixtures III: Autoregressive processes (*DAR*(*p*)). Naval Postgraduate School Technical Report NPS 55-78-022.

——— (1983) Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Series Anal.*, **4**, 18–36.

Johnson, N. L. and Kotz, S. (1969) *Distributions in Statistics: Discrete Distributions*. Boston: Houghton Mifflin.

Kanter, M. (1975) Autoregression for discrete processes mod 2. *J. Appl. Prob.*, **12**, 371–375.

Katz, L. and Proctor, C. H. (1959) The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika*, **24**, 317–327.

Katz, R. W. (1981) On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 243–249.

Kedem, B. (1980) *Binary Time Series*. New York: Marcel Dekker.

Klotz, J. (1973) Statistical Inference in Bernouilli trials with dependence. *Ann. Statist.*, **1**, 373–379.

Lloyd, E. H. (1977) Reservoirs with seasonally varying Markovian inflows and their first passage times. Research report RR-77-4, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Logan, J. A. (1981) A structural model of the higher-order Markov process incorporating reversion effects. *J. Math. Sociol.*, **8**, 75–89.

McKenzie, E. (1981) Extending the correlation structure of exponential autoregressive-moving-average processes. *J. Appl. Prob.*, **18**, 181–189.

Pegram, G. G. S. (1980) An autoregressive model for multilag Markov chains. *J. Appl. Prob.*, **17**, 350–362.
Raftery, A. E., Haslett, J. and McColl, E. (1982). Wind power: a space-time process? In *Time Series Analysis: Theory and Practice 2* (O. D. Anderson, ed.). Amsterdam: North-Holland.
Tong, H. (1975) Determination of the order of a Markov chain by Akaike's information criterion. *J. Appl. Prob.*, **12**, 488–497.

APPENDIX

Proofs of Results in Section 2

Proof of Theorem 1: Let S be the $m^l \times m^l$ transition matrix for the Markov chain with the m^l possible values of $(X_{t-1}, \ldots, X_{t-l})$ as states. The elements of S are

$$P\,[X_t = i_0, X_{t-1} = i_1, \ldots, X_{t-l+1} = i_{l-1} \mid X_{t-1} = j_1, \ldots, X_{t-l} = j_l]$$

$$= \begin{cases} \sum\limits_{k=1}^{l} \lambda_k q_{i_0 j_k}, \text{ if } i_k = j_k \text{ for } k = 1, 2, \ldots, l-1 \\ 0 \qquad\qquad\qquad\quad, \text{otherwise.} \end{cases}$$

Each row of S corresponds to an l-tuple $(i_0, \ldots, i_{l-1})$ and these are ordered in such a way that i_0 varies most slowly, i_1 second most slowly, and so on. Similarly, the columns of S corresponds to values of $(j_1, \ldots, j_l)$ with j_1 varying most slowly, and so on.

We first show that S is ergodic. All the states of S intercommunicate, and so S is irreducible. Amongst the diagonal elements of S, m are non-zero, corresponding to

$$i_0 = i_1 = \ldots = i_{l-1} = j_1 = \ldots = j_l = k \; (k = 1, \ldots, m),$$

as they have values $q_{kk} > 0$ by hypothesis. Thus these m states are aperiodic and hence since S is irreducible all its states are aperiodic. It follows by Cox and Miller (1965, p. 124) that S, being finite, specifies an ergodic Markov chain. Hence S has a unique equilibrium distribution ξ satisfying $S\xi = \xi$ with elements

$$\xi_{i_1, \ldots, i_l} = \lim_{t \to \infty} P\,[X_{t-1} = i_1, \ldots, X_{t-l} = i_l]$$

arranged so that i_1 varies most slowly, and so on. Let $\omega = (\omega_1, \ldots, \omega_m)'$ be the corresponding one-dimensional marginal equilibrium distribution. Also let R be the "collapsed form" of S as defined by Pegram (1980). This is an $m \times m^l$ matrix which consists of the elements of S not identically equal to zero. Clearly, in general

$$R\xi = \omega. \tag{A.1}$$

Also for the model (1.1)

$$R = \sum_{j=1}^{l} \lambda_j U_j$$

where $U_j = A_{j,1} X \ldots X A_{j,l}$ where

$$A_{j,k} = \begin{cases} Q \text{ if } j = k \\ 1' \text{ if } j \neq k \end{cases}$$

and X denotes the matrix right direct product $AXB = \{a_{ij}B\}$. We now calculate $R\xi$ in another way. The kth element of $U_j\,\xi$ is (where all summations are over $1, \ldots, m$)

$$\sum_{i_1, \ldots, i_l} q_{ki_j}\, \xi_{i_1, \ldots, i_l}$$

$$= \sum_{i_j} q_{ki_j} \sum_{\substack{i_h : h=1, \ldots, l \\ h \neq j}} \xi_{i_1, \ldots, i_l}$$

$$= \sum_{i_j} q_{ki_j} \, \omega_{i_j}$$

which is also the kth element of $Q\omega$. Thus

$$R\xi = \sum_{j=1}^{l} \lambda_j (Q\omega) = Q\omega \qquad \text{(A.2)}$$

Equating (A.1) and (A.2) shows that

$$Q\omega = \omega$$

to which the unique solution is $\omega = \pi$ by Cox and Miller (1965, p. 124).

Proof of Theorem 2. First consider the case where $k = 1, \ldots, l$. Let

$$S_{tk} = \{X_{t+k-g} : g = 1, \ldots, l, g \neq k\}$$

and $A_k = \{g : g = 1, \ldots, l, g \neq k\}$. Then

$$\begin{aligned} p_{ij}(k) &= \underset{S_{tk}}{E} \; \{P\,[X_{t+k} = i, X_t = j \mid S_{tk}]\} \\ &= \underset{S_{tk}}{E} \; \{P\,[X_{t+k} = i \mid X_t = j, S_{tk}] \; P\,[X_t = j \mid S_{tk}]\} \\ &= \sum_{g \in A_k} \lambda_g \underset{S_{tk}}{E} \; \{q_{i,\, X_{t+k-g}} P\,[X_t = j \mid S_{tk}]\} + \lambda_k \underset{S_{tk}}{E} \{q_{ij} P\,[X_t = j \mid S_{tk}]\} \qquad \text{(A.3)} \\ &= \sum_{g \in A_k} \lambda_g \underset{X_{t+k-g}}{E} \; \{q_{i,\, X_{t+k-g}} \; P\,[X_t = j \mid X_{t+k-g}]\} + \lambda_k q_{ij} \pi_j \\ &= \sum_{g \in A_k} \lambda_g \sum_{h=1}^{m} q_{ih} p_{hj}(k-g) + \lambda_k q_{ij} \pi_j \end{aligned}$$

But $q_{ij}\pi_j$ is the (i,j)th element of $QP(0)$, so this is the (i,j)th element of

$$\sum_{g=1}^{k} \lambda_g QP(k-g),$$

as required.

When $k = l+1, l+2, \ldots$ the result follows by an argument which is the same, except that the last term in (A.3) does not have to be considered separately. ■

Proof of the Corollary to Theorem 2. Let $\gamma = (\gamma_1, \ldots, \gamma_m)'$ where $\gamma_i = (i - E(X_t))/\sqrt{\text{var}(X_t)}$. Then $\rho_k = \gamma' P(k) \gamma$ and $\tilde{\rho}_k = \gamma' QP(k) \gamma$. The result then follows from (2.1). ■

Proof of Theorem 3. (i) The result holds provided the m^2 equations

$$P(1) = \lambda_1 QP(0) + (1 - \lambda_1) P(1)' \qquad \text{(A.4)}$$

in the m^2 unknowns $\{p_{ij}(1)\}$ have a unique solution. (A.4) may be written

$$p_{ij}(1) = \sum_{r,s=1}^{m} b_{ij;rs} \, p_{rs}(1) + c_{ij}, \qquad \text{(A.5)}$$

where

$$b_{ij;rs} = \begin{cases} (1-\lambda_1)\,q_{is}, & \text{if } j = r \\ 0 & \text{, otherwise} \end{cases}$$

and $c_{ij} = \lambda_1 q_{ij} \pi_j$. Then by Bellman (1970, p. 298), (A.5) has a unique solution of $\Sigma\, b_{ij;rs} < 1$ for $r, s = 1, \ldots, m$. But $\Sigma\, b_{ij;rs} = (1-\lambda_1) < 1$ by hypothesis, and so (A.5) has a unique solution, as required.

(iii) Let

$$\psi = 1 - \sum_{i=1}^{m} \eta_i.$$

Then Q can be written $Q = \psi U + \eta 1'$ where $U = \{u_{ij}\}$ is a non-negative matrix with column sums equal to one, so that U has all the properties of a transition matrix. (2.1) then becomes

$$P(k) = \sum_{j=1}^{l} \lambda_j Q P(k-j) \quad (k = 1, \ldots, l-1)$$

$$= \sum_{j=1}^{l} \lambda_j \psi U P(k-j) + \eta \pi' \qquad \text{(A.6)}$$

since $(\eta 1') P(h) = \eta \pi'$ $(h \in Z)$ and $\lambda_1 + \ldots + \lambda_l = 1$. Then, since $P(0) = \text{diag}\{\pi_1, \ldots, \pi_m\}$ and $P(-h) = P(h)'$ we can write (A.6) as

$$p_{ij}(k) = \sum_{v=1}^{l-1} \sum_{r,s=1}^{m} b_{ij;rs}(k,v)\, p_{rs}(v) + c_{ij}(k) \qquad \text{(A.7)}$$

where

$$b_{ij;rs}(k,v) = \begin{cases} \lambda_{k-v}\, \psi u_{ir}, & \text{if } j = s \text{ and } 1 \leqslant v \leqslant k-1 \\ \lambda_{k+v}\, \psi u_{is}, & \text{if } j = r \text{ and } 1 \leqslant v \leqslant l-k \\ 0 & \text{, otherwise} \end{cases}$$

and $c_{ij}(k) = \lambda_k \psi u_{ij} \pi_j + \eta_i \pi_j$.

By Bellman (1970, p. 298) (A.7) has a unique solution if

$$1 > \sum_{k=1}^{l-1} \sum_{i,j=1}^{m} b_{ij;rs}(k,v) \quad (r, s = 1, \ldots, m; v = 1, \ldots, l-1)$$

$$= \sum_{k=v+1}^{l-1} \sum_{i=1}^{m} \lambda_{k-v} \psi u_{ir} + \sum_{k=1}^{l-v} \sum_{i=1}^{m} \lambda_{k+v} \psi u_{is}$$

$$= \psi \left(\sum_{k=1}^{l-1-v} \lambda_k + \sum_{k=v+1}^{l} \lambda_k \right)$$

since

$$\sum_{i=1}^{m} u_{ir} = 1 \ (r = 1, \ldots, m).$$

Since $\lambda_k \geqslant 0$ by hypothesis this is greatest when $v = 1$, in which case it is equal to

$$\psi(2 - \lambda_1 - \lambda_{l-1} - \lambda_l) < 1$$

by hypothesis (2.3). Thus (A.7) has a unique solution, as required.

(ii) Clearly (iii) holds when $l = 3$, in which case $(2 - \lambda_1 - \lambda_{l-1} - \lambda_l) = 1$. Also, suppose the ith row has no zero elements. Then $\eta_i > 0$ and so $\psi \leqslant 1 - \eta_1 \leqslant 1$ so that (2.3) is satisfied. ■