

MARKOV CHAIN MODELS FOR HYDROLOGIC TIME SERIES ANALYSIS

Felino P. Lansigan^{1/}

1. Introduction

In recent years, several stochastic models have been proposed for modelling time series which are claimed to have the capability to reproduce the main statistical properties of the observed historical series. Some models which have gained increased popularity in time series analysis in general and in hydrology and water resources in particular are the mixed autoregressive-moving average models of order (p,q) or ARMA (p, q) models (Box and Jenkins, 1970) and the Markov chain models (e.g., Lloyd, 1967).

One immediate problem in using Markov chain models in modelling of continuous time series involves the formulation of appropriate transition probability matrix models that account for the correlation structure in the sequence. For instance, in those cases where the autoregressive models or the mixed autoregressive-moving average models are used to represent the dependence structure of the series, the main problem is that of determining the transition probabilities of the corresponding Markov chain models that will mimic the behavior of the series. Some other modelling activities, on the other hand, may require that modelling of annual time series maintain and take into account the seasonal structure of the associated seasonal series when aggregated to form the annual sequence.

^{1/}Assistant Professor in Statistics, Statistical Laboratory, Institute of Mathematical Sciences and Physics, U.P. at Los Banos, College, Laguna.

This paper presents some transition probability matrix models in Markov chain modelling of hydrologic time series. Their properties such as their correlation structure and resemblance with their continuous model counterparts are analyzed. The structure of the aggregate of seasonal series modelled by Markov chains is also examined. Some practical applications are also discussed.

2. Markov Chain Modelling

The discrete modelling of time series using Markov chains involves two important steps, namely, the definition of states of the Markov chain and the estimation of the transition probabilities. Basic to these steps is the twin problem of the nature of discretization and the size of the unit of discretization used. The formulation of an explicit transition matrix model is the main subject of this paper.

2.1 Estimation of Transition Matrix

Most annual hydrologic time series have dependence structures which can be modelled by an autoregressive model of order one, or the AR (1) model. Markov chain modelling of hydrologic time series with an AR (1) dependence structure requires efficient estimation of the transition probability matrix of the corresponding Markov chain model. A logical approach to follow is to utilize the transition probability matrix derived from a discrete analogue of the continuous autoregressive model.

In order to represent the dependence structure of the time series by a Markov chain, Pegram (1972) developed a model which

defines the transition matrix explicitly. He defined the transition model P of the stochastic process $\{X_t\}$ as

$$P = \rho I + (1-\rho) \underline{\mu} \mathbf{1}^T \quad (1)$$

$$0 < \rho < 1, \underline{\mu} \geq \underline{0}, \mathbf{1}^T \underline{\mu} = 1$$

where $\mathbf{1}^T$ denotes the vector $(1 \ 1 \ \dots \ 1)$ of size $i \times k$ where k is the number of discrete intervals, $\underline{\mu}$ is a column vector of the stationary distribution of $\{X_t\}$, I is the identity matrix of size $k \times k$, and ρ is the lag-one correlation coefficient of $\{X_t\}$.

The limiting stationary distribution of the transition matrix is the $\underline{\mu}$ vector itself. Thus, to fit the model of eq. (1), the only parameters that have to be estimated are the stationary distribution $\underline{\mu}$ and the lag-one serial correlation coefficient ρ . Therefore, it is possible to specify any prior stationary distribution and correlation ρ and then compute the corresponding transition matrix.

It may be noted that the autocorrelation structure of $\{X_t\}$ is given by (Pegram, 1972)

$$\text{corr}(X_t, X_{t+k}) = \rho^k, \quad k = 1, 2, 3, \dots \quad (2)$$

which is identical to that of an AR (1) model defined as

$$X_t = \phi X_{t-1} + \epsilon_t$$

where ϕ is the autoregressive coefficient which can be shown to be equal to the first-order autocorrelation coefficient ρ and ϵ_t is the independent residual (Box and Jenkins, 1970). Moreover, since eq. (2) is independent of the size of the state space of $\{X_t\}$ it is possible to increase the size of the state space without changing the correlation structure of $\{X_t\}$. This is an appealing property especially when

representing a continuous process by a discrete model.

2.2 Mixture of Independent Markov Chains

2.2.1 First-Order Markov Process

Let $\{X_t\}$ be a homogeneous simple Markov chain with transition matrix P_X defined as in eq. (1) by

$$P_X = \rho_X I + (1-\rho_X) \underline{\mu}_X \mathbf{1}^T$$

Likewise, let $\{Y_t\}$ be another simple Markov chain with transition matrix P_Y given by

$$P_Y = \rho_Y I + (1-\rho_Y) \underline{\mu}_Y \mathbf{1}^T$$

Furthermore, assume that $\{X_t\}$ and $\{Y_t\}$ are independent among themselves. It can be shown that $\{X_t\}$ and $\{Y_t\}$ each have a correlation structure identical to an AR (1) process; i.e.,

$$\text{corr}(X_t, X_{t+k}) = \rho_X^k, \quad k = 1, 2, 3, \dots$$

and

$$\text{corr}(Y_t, Y_{t+k}) = \rho_Y^k, \quad k = 1, 2, 3, \dots$$

The process $\{Z_t\}$ where $Z_t = X_t + Y_t$ is a mixture of X_t and Y_t . The marginal distribution of Z_t can be determined as follows. Let the state space of $\{X_t\}$ be $\{-m_X, -(m_X - 1), \dots, -1, 0, +1, \dots, (m_X - 1), m_X\}$

where

$$P(X_t = i) = P_X(i), \quad -m_X \leq i \leq m_X.$$

Similarly, let the state space of $\{Y_t\}$ be $\{-m_Y, -(m_Y - 1), \dots, -1, 0, +1, \dots, (m_Y - 1), m_Y\}$ where

$$P(Y_t = i) = P_Y(i) \quad -m_Y \leq i \leq +m_Y$$

Since X_i and Y_i are independent, then

$$P(Z_t = k) = \sum_i \rho_X(i) \rho_Y(k-i),$$

$$-(m_X + m_Y) \leq k \leq + (m_X + m_Y)$$

where $-m_X \leq i \leq + m_X$ and such that $-m_Y \leq k-i \leq + m_Y$.

By definition, the mean and variance of $\{Z_t\}$ is given by

$$E(Z_t) = E(X_t) + E(Y_t)$$

and

$$\text{Var}(Z_t) = \text{Var}(X_t) + \text{Var}(Y_t)$$

The autocorrelation function of the $\{Z_t\}$ process is given by

$$\text{corr}(Z_t, Z_{t+k}) = \frac{\rho_X^k \sigma_X^2 + \rho_Y^k \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}, \quad k = 1, 2, \dots \quad (3)$$

It may be noted that the autocorrelation function of $\{Z\}$ is a weighted linear combination of ρ_X^k and ρ_Y^k where the weights are determined by the individual variances. Thus, when $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, then eq. (3) reduces to

$$\text{corr}(Z_t, Z_{t+k}) = \frac{\rho_X^k + \rho_Y^k}{2}, \quad k = 1, 2, \dots$$

In the particular case $\rho_X = \rho_Y = \rho$, eq. (3) yields

$$\text{corr}(Z_t, Z_{t+k}) = \rho^k, \quad k = 1, 2, \dots$$

Thus, in this case the dependence structure of $\{Z_t\}$ is identical to the autocorrelation structure of each individual Markov chain.

Note that the $\{Z_t\}$ process itself is not a Markov chain since $P(Z_t = i | Z_{t-1} = j, Z_{t-2} = k) \neq P(Z_t = i | Z_{t-1} = j)$, in general.

However, consider now the states of the mixture process to be the pair (X_t, Y_t) .

Since X_t and Y_t are independent, and each

individual process has Markovian structure, then,

$$P(X_t = i, Y_t = j | X_{t-1} = k,$$

$$Y_{t-1} = \ell, X_{t-2} = m, Y_{t-2} = n)$$

$$= P(X_t = i, Y_t = j | X_{t-1} = k, Y_{t-1} = \ell)$$

Thus, the pair $\{X_t, Y_t\}$ forms a bivariate Markov chain whose transition probabilities are given by

$$a(ij|k\ell) = P_X(i, k) P_Y(j, \ell)$$

where

$$a(ij|k\ell) = P(X_t = i, Y_t = j | X_{t-1} = k,$$

$$Y_{t-1} = \ell), P_X(i, k) = P(X_t = i | X_{t-1} = k),$$

$$\text{and } P_Y(j, \ell) = P(Y_t = j | Y_{t-1} = \ell).$$

2.2.2 ARMA (1,1) Process

Consider again the $\{X_t\}$ and $\{Y_t\}$ processes discussed earlier. Define a mixture process $\{X_t^*\}$ such that

$$X_t^* = X_t + U_t$$

where $\{X_t\}$ is a simple Markov chain and $\{U_t\}$ is an independent discrete white noise process with mean zero and variance σ^2 .

As before, let the state space of $\{X_t\}$ be $\{-m_X, (-m_X-1), \dots, -1, 0, +1, \dots, (m_X-1), m_X\}$

where

$$P(X_t = i) = P_X(i), \quad -m_X \leq i \leq m_X$$

Since X_t and U_t are independent, the marginal probability distribution of X_t^* is given by

$$P(X_t^* = k) = \sum_i P_X(i) U(k-i),$$

$$-m_X - a \leq k \leq m_X + a$$

where

$-m_x \leq i \leq m_x$ and such that $-a \leq k - i \leq a$.

The mean and the variance of X_t^* are given by

$$E(X_t^*) = E(X_t) + E(U_t)$$

and

$$\text{Var}(X_t^*) = \text{Var}(X_t) + \text{Var}(U_t)$$

The mixture process $\{X_t^*\}$ will be correlated due to the dependence in the $\{X_t\}$ process. It can be shown that the correlation structure of $\{X_t^*\}$ is given by

$$\text{corr}(X_t^*, X_{t+k}^*) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \rho_x^k$$

$k = 1, 2, \dots$ (4)

Notice that eq. (4) is identical to the correlation structure of an ARMA (1,1) process (Box and Jenkins, 1970).

Note the autocorrelation structure of $\{X_t^*\}$ is independent of the state space. Hence, the state space can be increased without altering the dependence structure of the process.

Note further that $\{X_t^*\}$ is not a simple Markov chain. However, if the mixture process $\{X_t^*\}$ is represented in terms of the pair (X_t, U_t) then the sequence $\{X_t, U_t\}$ forms a bivariate Markov chain with one-step transition probabilities defined by

$$c(ij|k\ell) = P_X(i,k) \cdot u(j)$$

Where

$$c(ij|k\ell) = P(X_t=i, U_t=j | X_{t-1}=k,$$

$$Y_{t-1}=\ell), P_X(i,k) = P(X_t=i | X_{t-1}=k),$$

Define now a mixture process $\{Z_t^*\}$ where

$$Z_t^* = X_t^* + Y_t^* = X_t + U_t + Y_t + W_t$$

It can be shown that the marginal probability distribution of Z_t^* is given by

$$P(Z_t^* = k) = \sum_i \sum_j \sum_\ell P_X(j) \cdot u(k-j) \cdot P_Y(\ell) \cdot w(k-i|\ell)$$

where $-(m_x + m_y + a + b) \leq i \leq (m_x + m_y + a + b)$, $-m_x \leq j \leq m_x$, $-a \leq k - j \leq a$, $-m_y \leq \ell \leq m_y$, and $-b \leq k - i - \ell \leq b$ with m_x and a , and m_y and b corresponding to X^* and Y^* , respectively.

The mean and variance of Z_t^* are given by

$$E(Z_t^*) = E(X_t) + E(Y_t)$$

and

$$\text{Var}(Z_t^*) = \text{Var}(X_t) + \text{Var}(Y_t) + \text{Var}(U_t) + \text{Var}(W_t).$$

The correlation structure of Z_t^* is defined by

$$\text{corr}(Z_t^*, Z_{t+k}^*) = \frac{\sigma_x^2 \rho_x^k + \sigma_y^2 \rho_y^k}{\sigma_x^2 + \sigma_u^2 + \sigma_y^2 + \sigma_w^2} \quad (5)$$

$(k = 1, 2, \dots)$

In the special case $\rho_x = \rho_y = \rho$, eq. (5) yields

$$\text{corr}(Z_t^*, Z_{t+k}^*) = \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2 + \sigma_u^2 + \sigma_y^2 + \sigma_w^2} \rho^k$$

$k = 1, 2, \dots$

which is identical to the dependence structure of an ARMA (1,1) process.

The mixture process $\{Z_t\}$ does not form a Markov chain. However, it can be shown that the quadruple $\{X_t, U_t, Y_t, W_t\}$ forms a quadrivariate Markov chain with one-step transition probabilities defined by

$$q(ijk\ell|efgh) = p_X(i,e) \cdot u(j)$$

$$p_Y(k,g) \cdot w(\ell)$$

where $q(ijk\ell|efgh) = p(X_t = i, U_t = j,$

$$Y_t = k, W_t = \ell | X_{t-1} = e, U_{t-1} = f,$$

$$Y_{t-1} = g, W_{t-1} = h), \text{ and}$$

$$p_X(\dots), p_Y(\dots), u(\cdot) \text{ and } w(\cdot)$$

are as defined earlier.

2.3 Aggregate of Seasonal Markov Chains

The annual sequence is the aggregate of the seasonal series. In discrete modelling of time series, it may be of particular interest to determine the model structure of the resulting annual series knowing the structure of the corresponding seasonal series.

Thus, let $\{X_{t,\tau}\}$, $t = 1, 2, \dots$ and $\tau = 1, 2, \dots, w$ (number of seasons), be a Markov chain with seasonal transition matrix defined for each season where t denotes the year and τ denotes the season. Let Z_t be a discrete process representing the aggregation of the seasonal values defined by $Z_t = \sum_{\tau=1}^w X_{t,\tau}$. The mean and variance of Z_t are given by

$$E(Z_t) = \sum_{\tau=1}^w \mu_\tau$$

and

$$\text{Var}(Z_t) = \sum_{\tau=1}^w \sigma_\tau^2 + \sum_{\tau=1}^w \sum_{\tau'=\tau+1}^w \rho_{\tau,\tau'} \sigma_\tau \sigma_{\tau'}$$

where μ_τ and σ_τ^2 are the mean and variance, respectively, for season τ , and $\rho_{\tau,\tau'} = \text{corr}(X_{t,\tau}, X_{t,\tau'})$.

The correlation structure of Z_t is given, in general, by

$$\text{corr}(Z_t, Z_{t+k}) = \frac{\sum_{\tau=1}^w \rho_\tau(k) \sigma_\tau^2 + \sum_{\tau=1}^w \sum_{\tau'=1}^w \rho_{\tau,\tau'}(k) \sigma_\tau \sigma_{\tau'}}{\sum_{\tau=1}^w \sigma_\tau^2 + 2 \sum_{\tau < \tau'} \rho_{\tau,\tau'} \sigma_\tau \sigma_{\tau'}}$$

$$k = 1, 2, \dots$$

where $\rho_\tau(k)$ is the correlation function for season τ , and $\rho_{\tau,\tau'}(k)$ is the lag- k correlation between seasons τ and τ' .

It may be noted that the aggregated process $\{Z_t\}$ does not form a Markov chain since $P(Z_t = i | Z_{t-1} = j, Z_{t-2} = k) \neq P(Z_t = i | Z_{t-1} = j)$ in general. However, if the aggregation is represented by the multivariate process $\{Z_t, X_{t,1}, Z_{t,2}, \dots, X_{t,w}\}$, a multivariate Markov chain is now formed with one-step multivariate transition matrix defined by

$$P(Z_{t+1} = v, X_{t+1,1} = r_1, X_{t+1,2} = r_2, \dots,$$

$$X_{t+1,w} = r_w | Z_t = u, X_{t,1} = s_1, X_{t,2}$$

$$s_2, \dots, X_{t,w} = s_w)$$

$$= \begin{cases} 0 & \text{if either } \sum_{i=1}^w r_i \neq v \text{ or } \sum_{i=1}^w s_i \neq u \\ P(X_{t+1,1} = r_1, X_{t+1,2} = r_2, \dots, X_{t+1,w} = r_w | \\ X_{t,1} = s_1, X_{t,2} = s_2, \dots, X_{t,w} = s_w) & \text{otherwise.} \end{cases}$$

Note that although the Markovian structure of the aggregation of the seasonal series has been retained, it is achieved at the expense of increasing the dimensionality of the state of the system.

3. Application of the Models

Time series modelling is mainly used in hydrology and water resources for generation of synthetic hydrologic sequences and for forecasting future hydrologic time series that reproduce the main statistical characteristics of the historical hydrologic time series. However, some other statistical properties of hydrologic series essential in water resources development studies are also important to look at. For instance, the range and the deficit of cumulative (partial) sums are properties related to storage capacities of reservoirs. Gomide (1975) has shown that range and deficit analysis follows directly from the theory of Markov chains.

The sequence of binomial net inflows into a reservoir with Markovian dependence structure can be represented by a Markov chain with transition matrix defined by eq. (1). The mean and the variance of range R_n and of deficit D_n for sample size n for various values of correlation coefficients ρ and size of state space m following Gomide (1975), and Lansigan (1982) are given in Table 1.

Another application of the Markov chain models presented in this paper involves the development of flooding models for two independent rivers passing through a junction. For instance, the streamflow downstream of the junction is the sum of water flowing from each river which can be modelled as two separate Markov chains. The

properties of the aggregated flows at the junction may be determined by considering the properties of the individual Markov model.

4. Summary and Conclusion

Discrete modelling of continuous time series using Markov chains requires formulation of adequate transition matrix models which have the same dependence structure as the sequence. Some explicit transition matrix models with correlation structures identical to that of the ARMA models are presented. The dependence structure of the aggregate of seasonal Markovian sequence is also examined.

Further research is needed to develop a family of transition matrix models analogous to the theoretical probability distributions available for curve fitting. Results presented in this paper suggest a possible strategy or approach for developing such models.

REFERENCES

- Box, G.E.P. and G. Jenkins. 1970. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
- Gomide, F. L. S. 1975. Range and Deficit Analysis Using Markov Chains, Hydrology Paper 79, Colorado State University, Fort Collins, Colorado.
- Lansigan, F. P. 1982. Storage Analysis for Annual and Seasonal Flows. Ph.D. Thesis, Colorado State University, Fort Collins, Colorado.
- Lloyd, E.H. 1967. Stochastic Reservoir Theory *Advances in Hydrosience*, Vol. 4, pp. 281-339.
- Pegram, G. G. S. 1972. Some Applications of Stochastic Reservoir Theory. Ph.D. Thesis, University of Lancaster, U.K.

Table 1. Distribution properties of R_n and D_n of binomial distributed net inputs X_t with $P(X_t = i) = \binom{2m}{m+i} (1/2)^{2m}$,
 $i = -m, -m+1, \dots, 0, \dots, m-1, m$.

ρ	m	$n = 2$	$n = 4$	$n = 8$	$n = 16$	$n = 32$				
0	∞	1.3621	2.2217	3.4879	5.3171	7.9322	$E(R_n)$			
	∞	.80	1.44	2.46	3.93	5.96	$E(D_n)^*$			
0	18	1.3546	2.2124	3.4772			$E(R_n)$			
	8	1.3453	2.2010	3.4640	5.2908					
	2	1.2969	2.1408	3.3943	5.2139	7.8220				
	18	.7790	1.0271	1.3976			$\sqrt{\text{Var}(R_n)}$			
	8	.7847	1.0312	1.4008	1.9408					
	2	.8136	1.0518	1.4178	1.9540	2.7266				
	0	18	.7924	1.4648	2.4740			$E(D_n)$		
		8	.7855	1.4552	2.4624	3.9015				
		2	.7500	1.4079	2.4026	3.8309	5.8751			
		18	.8283	1.1200	1.5067			$\sqrt{\text{Var}(D_n)}$		
		8	.8315	1.1234	1.5089	2.0828				
		2	.8478	1.1419	1.5199	2.0916	2.9245			
03		8	1.4130	2.4599				$E(R_n)$		
		2	1.3578	2.3969	4.0243	6.4411	9.9586			
		8	.9493	1.4026						
		2	.9984	1.4399	1.9945	2.7331	3.7549	$\sqrt{\text{Var}(R_n)}$		
		03	8	.7855	1.5118				$E(D_n)$	
			2	.7500	1.4534	2.6651	4.5418	7.2850		
	8		.9480	1.4411						
	2		.9666	1.4678	2.0881	2.8957	4.0360	$\sqrt{\text{Var}(D_n)}$		
	05		2	1.3984	2.5679	4.5016	7.4821	11.8771	$E(R_n)$	
			2	1.1028	1.7417	2.5452	3.5460	4.8993	$\sqrt{\text{Var}(R_n)}$	
			05	2	.7500	1.4762	2.8130	5.0520	8.4616	$E(D_n)$
				2	1.0383	1.7088	2.6010	3.7082	5.1742	$\sqrt{\text{Var}(D_n)}$
07			2	1.4391	2.7396	5.0404	8.8276	14.6463	$E(R_n)$	
			2	1.1967	2.0785	3.3424	4.9178	6.8719	$\sqrt{\text{Var}(R_n)}$	
			07	2	.7500	1.4914	2.9265	5.5536	9.9288	$E(D_n)$
				2	1.1054	1.9697	3.2843	5.0244	7.1832	$\sqrt{\text{Var}(D_n)}$

*Note: Approximate values from Fig. 5.9, Gomide (1975); values of $E(R_n)$ and $\sqrt{\text{Var}(R_n)}$ from Pegram et.al., (1980); values of $E(D_n)$ and $\sqrt{\text{Var}(D_n)}$ from Lansigan (1982).