

Chapter 2

Markov Chains and Their Extensions

Abstract This chapter deals with likelihood-based inference for ergodic finite as well as infinite Markov chains. We also consider extensions of Markov chain models, such as Hidden Markov chain, Markov chains based on polytomous regression, and Raftery's Mixture Transition Density model. These models have less number of parameters as compared to a higher order finite Markov chain. Lastly, we discuss methods of estimation in grouped data from finite Markov chains.

2.1 Markov Chains

Let $\mathfrak{X} = (X_0, X_1, \dots)$ be an M -state Markov chain with the state-space $S = \{1, \dots, M\}$ and the one-step transition probability matrix (t.p.m.) $P = ((p_{ij}))_{M \times M}$, $p_{ij} \geq 0, \forall (i, j)$ and $\sum_{j=1}^M p_{ij} = 1, \forall i$. We assume that all states communicate with each other and are aperiodic, which implies that all states are non-null persistent and hence the chain is ergodic. Let $\{\pi_i, i = 1, 2, \dots, M\}$ be the unique stationary distribution of the Markov chain $\mathfrak{X}$. It then follows that for all i, j ,

$$p_{ij}^{(t)} \rightarrow \pi_j > 0, \text{ as } t \rightarrow \infty.$$

Let $(X_0, X_1, \dots, X_T)$ be the $T + 1$ successive observations from the above Markov chain. The conditional likelihood (given the initial observation X_0) is given by $L(P) = \prod_{t=0}^T p_{x_t, x_{t+1}}$. Thus,

$$\ln(L(P)) = \sum_i \sum_j N_{ij} \ln p_{ij} = \sum_{t=0}^{T-1} I[X_t = i, X_{t+1} = j].$$

We need to maximize $\ln(L(P))$ with respect to p_{ij} 's, subject to the constraints that $\sum_j p_{ij} = 1, \forall i$. Let λ_i 's be the Lagrangian parameters. We set

$$f = \ln(L(P)) + \sum_{i=1}^M \lambda_i \left(\sum_j p_{ij} - 1 \right).$$

Then, if $p_{ij} > 0$, it can be easily shown that

$$\hat{p}_{ij} = \frac{N_{ij}}{N_{i+}}, \quad \text{where } N_{i+} = \sum_j N_{ij}.$$

Let $Z_t = I[X_t = i, X_{t+1} = j]$. We note that Z_t is a Bernoulli random variable. Let $\{p_i, i = 1, 2, \dots, M\}$ be the initial distribution of $\mathfrak{X}$. Then,

$$E[N_{ij}] = \sum_{t=0}^{T-1} E[Z_t] = \sum_{t=1}^{T-1} \sum_{k=1}^M \left(p_k p_{ki}^{(t)} p_{ij} \right).$$

Since $p_{ki}^{(t)} \rightarrow \pi_i$, $\sum_{t=1}^{T-1} p_{ki}^{(t)} / T \rightarrow \pi_i$. Thus, $E[N_{ij}/T] \rightarrow \pi_i p_{ij}$. Further, we observe that

$$\text{Var}[N_{ij}] = \sum_{t=1}^{T-1} \sum_{k=1}^M \left(p_k p_{ki}^{(t)} p_{ij} \right) \left(1 - \left(p_k p_{ki}^{(t)} p_{ij} \right) + \sum_{k=1}^M \sum_{s \neq t} \left(p_k p_{ki}^{(s)} p_{ij} \right) p_{ki}^{(t)} p_{ij} \right)$$

and that $V[N_{ij}/T] \rightarrow 0$ as $T \rightarrow \infty$. It follows that, for all i, j , $N_{ij}/T \rightarrow \pi_i p_{ij}$ in probability and hence $\hat{p}_{ij} \rightarrow p_{ij}$ in probability. We observe that the above convergence holds for any initial distribution $\{p_i, i = 1, 2, \dots, M\}$. It is easy to derive the Fisher information matrix from the above likelihood. It follows from Theorem 1.5.1 (and also from Theorem 1.1 of Billingsley 1961) that the joint distribution of $\sqrt{T}(\hat{p}_{ij} - p_{ij})$ $i, j \in S$ (written in a suitable vector form) is asymptotically multivariate normal with the mean vector 0 and variances and covariances given by

$$V(\sqrt{T}(\hat{p}_{ij} - p_{ij})) \approx \pi_i p_{ij}(1 - p_{ij})$$

$$\text{Cov}(\sqrt{T}(\hat{p}_{ij} - p_{ij}), \sqrt{T}(\hat{p}_{ik} - p_{ik})) \approx -\pi_i p_{ij} p_{ik}, \quad j \neq k.$$

$$\text{Cov}(\sqrt{T}(\hat{p}_{ij} - p_{ij}), \sqrt{T}(\hat{p}_{lk} - p_{lk})) \approx 0, \quad \text{if } i \neq l.$$

It may be remarked that the likelihood function and hence the variance-covariance matrix resemble the likelihood function and variance-covariance pattern of M independent multinomial distributions respectively.

Goodness of fit of a Markov chain

1. LRT and Pearson's X^2 statistics for testing the order. The first order Markov property can be judged by testing against the second order. Under the assumption of the second order, the ML estimate of $p_{ij,k} = P[X_{t+2} = k | X_t = i, X_{t+1} = j]$ is

given by $\hat{p}_{ij,k} = \frac{N_{ijk}}{N_{ij+}}$, where $N_{ijk} = \sum_{t=1}^{T-2} I[X_t = i, X_{t+1} = j, X_{t+2} = k]$ and $N_{ij+} = \sum_k N_{ijk}$. We assume that the vector valued Markov chain $\{(X_t, X_{t+1}), t \geq 0\}$ is irreducible and aperiodic. Thus, the LRT statistic is given by

$$2 \left\{ \sum_{ijk} N_{ijk} \ln \hat{p}_{ijk} - \sum_{ij} N_{ij} \ln \hat{p}_{ij} \right\} \sim \chi_{(M-1)^2 M}^2,$$

as the difference in the numbers of parameters of the two models is $M^2(M-1) - M(M-1) = (M-1)^2 M$. The corresponding Pearson's X^2 -statistic is given by

$$X^2 = \sum_{ijk} \frac{(N_{ijk} - E_{ijk})^2}{E_{ijk}} \text{ where } E_{ijk} = N_{i++} \hat{p}_{ij} \hat{p}_{jk}.$$

The degrees of freedom are the same as that of the LRT. One can similarly obtain ML estimates of an L order Markov chain and develop a large sample chi-squared test for an L -th order Markov chains against a Markov chain of order $(L+1)$.

2. Estimation of the true order of a Markov chain (Selection of a Markov model). The above method of testing an order L against $L+1$ is sequential in nature. We continue to test until an r -order model is not rejected against the $(r+1)$ -th order model. Since this procedure involves a series of tests to be carried out in a sequential manner, the probability of selecting the underlying true model (of unknown order) may not asymptotically converge to 1. A better and theoretically valid procedure is to apply the information criteria such as akaike's information criterion (AIC) or bayes information criterion (BIC). These are defined below. Let K be the number of parameters of a model under consideration.

$$\text{AIC} = 2 \sup_{\theta} \ln L(\theta) - T,$$

$$\text{BIC} = 2 \sup_{\theta} \ln L(\theta) - K \ln(T).$$

Both the log-likelihood and the number K of parameters change from model to model. We select the model which has the smallest AIC/BIC. Katz (1981) has shown that the BIC procedure gives a consistent estimator of the true order of a Markov model. The AIC frequently overestimates the true order. Another advantage of both AIC and BIC is that they can be applied even if the models are not nested. A construction of LRT requires the nested property, e.g., the L -th order Markov model is included in the $(L+1)$ -th order Markov model.

3. Time homogeneity. The assumption of time homogeneity can be tested by dividing the data into several non-overlapping blocks of consecutive observations of moderate length. We then compute $\hat{p}_{ij}$ for each (i, j) obtained from each of such blocks and plot the estimators against the time. If the transition probabilities vary a lot over

time, such plots would reveal changes. (A formal test is difficult to construct.) The procedure can be carried out based on overlapping blocks also.

4. *Distribution of patterns.* Application of a two-state Markov chain to a sequence of dry (the state 1) and wet (rainy) (the state 0) days over a monsoon at a place requires fitting of distributions of lengths of wet and dry cycles or spells. Length of a wet cycle equals r with probability $p_{00}^{r-1}p_{01}$. Observed distribution of wet cycle lengths can be compared with the distribution with estimated parameters. Visual inspection of plots suffices in most of the cases. The Pearson's X^2 statistic can be computed while comparing with other models. It may be pointed out that the X^2 statistic does not have a chi-square distribution. We refer to Table 2.10 (p. 76) of Guttorp (1995) for such an application.

2.2 Parametric Models

A parametric model with M states for $\mathfrak{X}$ is given by specifying p_{ij} as functions of θ , a p -dimensional parameter, i.e., as $p_{ij}(\theta) = P_\theta[X_{t+1} = j | X_t = i]$. The parameter space is Θ , an open subset of p -dimensional Euclidean space. We make the following assumptions.

1. The set $D = \{(i, j) | p_{ij}(\theta) > 0\}$ does not depend upon θ .
2. Each of the functions $p_{ij}(\theta)$ is twice differentiable with respect to θ_r , $r = 1, 2, \dots, p$. The second derivatives are continuous.
3. The appropriately constructed matrix $\frac{\partial p_{ij}(\theta)}{\partial \theta_r}$, $(i, j) \in D$, $r = 1, 2, \dots, p$ of order $d \times p$ is of rank p where d is the number of elements in the set D .
4. $\forall \theta$, $\mathfrak{X}$ is an irreducible, non-null persistent, aperiodic Markov chain.

The above assumptions can be described as extensions of Cramer regularity conditions, assumed usually in the case of i.i.d. observations. Let $\pi_i(\theta)$ denote the unique stationary distribution of $\mathfrak{X}$. For the time being, we assume that the chain is in equilibrium, i.e., X_0 follows the distribution $\pi_i(\theta)$. It is easily seen that $\ln L(\theta) = \sum_{i,j} N_{ij} \ln p_{ij}(\theta) + \sum_i I[X_0 = i] \pi_i(\theta)$. We ignore the second term as before. Consequently, the likelihood equations are given by

$$\frac{\partial \ln L(\theta)}{\partial \theta_r} = \sum_{(i,j) \in D} \frac{\partial p_{ij}(\theta)}{\partial \theta_r} \frac{N_{ij}}{p_{ij}(\theta)} = 0, \quad r = 1, 2, \dots, p.$$

Further,

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta_r \partial \theta_s} = \sum_{(i,j) \in D} \left[\frac{N_{ij}}{p_{ij}(\theta)} \frac{\partial^2 p_{ij}(\theta)}{\partial \theta_r \partial \theta_s} - \frac{N_{ij}}{(p_{ij}(\theta))^2} \frac{\partial p_{ij}(\theta)}{\partial \theta_r} \frac{\partial p_{ij}(\theta)}{\partial \theta_s} \right].$$

Now, since $N_{ij} = \sum_{t=0}^{T-1} I[X_t = i, X_{t+1} = j]$, in view of stationarity, we have

$$E(N_{ij}) = T\pi_i(\theta)p_{ij}(\theta).$$

Since $\sum_j p_{ij}(\theta) = 1$, $\sum_{(i,j) \in D} \pi_i \frac{\partial^2 p_{ij}(\theta)}{\partial \theta_r \partial \theta_s} = 0$ in view of the regularity conditions. The Fisher Information Matrix is defined by

$$I(\theta) = \left(\left(\lim_{T \rightarrow \infty} \frac{1}{T} E \left[\frac{\partial \ln L(\theta)}{\partial \theta_r} \frac{\partial \ln L(\theta)}{\partial \theta_s} \right] \right) \right).$$

In view of the regularity conditions, it is then given by

$$I(\theta) = \left((I_{rs}(\theta)) \right),$$

where

$$I_{rs}(\theta) = - \sum_{(i,j) \in D} \frac{\pi_i(\theta)}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_r} \frac{\partial p_{ij}(\theta)}{\partial \theta_s}.$$

The expected values in the above expressions have been derived with respect to the joint distribution of (X_0, X_1) under the assumption of stationarity. Let $\hat{\theta}$ be a consistent solution of the likelihood equations. It follows that

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N_p \left(\mathbf{0}, (I(\theta))^{-1} \right).$$

An estimator of the Fisher Information matrix is needed in construction of confidence regions and tests of hypotheses. In practice, it is easier to use the observed Fisher Information matrix, whose (i, j) -th element is given by

$$F_{ij}(\hat{\theta}) = - \left. \frac{\partial^2 \ln L(\theta)}{\partial \theta_r \partial \theta_s} \right|_{\theta=\hat{\theta}}$$

and the corresponding estimator of $I(\theta)$ is given by

$$\hat{I}(\theta) = \left((F_{ij}(\hat{\theta})/T) \right).$$

Another estimator of $I(\theta)$ is $I(\hat{\theta})$, which is obtained by replacing θ by $\hat{\theta}$ in $I(\theta)$. However, this requires a theoretical derivation of the expectations involved, which could be tedious in some cases.

Goodness of fit of parametric finite Markov chain. models: This is similar to goodness-of-fit procedures for parametric multinomial models. We note that under H_0 , $\mathcal{X}$ follows the above parametric model, $E_{H_0}(N_{ij}) = T\pi_i(\theta)p_{ij}(\theta)$, the estimate of which is given by $E_{ij} = N_{i+}p_{ij}(\hat{\theta})$, under H_0 . Thus, the Pearsonian X^2 statistic is given by

$$X^2 = \sum_{(i,j) \in D} \frac{(N_{ij} - N_{i+}p_{ij}(\hat{\theta}))^2}{N_{i+}p_{ij}(\hat{\theta})}$$

and its asymptotic null distribution is χ^2 with (the number of elements in the set D) – $M - p$ as the degrees of freedom. The LRT statistic has the same asymptotic distribution, under H_0 .

One can consider the Q–Q plot of $(N_{ij} - N_{i+}p_{ij}(\hat{\theta})) / \sqrt{N_{i+}p_{ij}(\hat{\theta})}$ for $(i, j) \in D$ by regrading them observations from $N(0, 1)$. Such a plot may reveal cells which have a sizable contribute to the LRT or X^2 , cf. Davison (2003), p. 235.

Testing for sub-models. Under the above stated conditions, we also get the chi-square distribution of the LRT for the null hypothesis $H_0 : \theta \in \Theta_0$, provided that the regularity conditions hold for the parameter space Θ_0 . The degrees of freedom for the chi-square are given by the $p - p_0$, where p_0 is the number of (distinct) parameters corresponding to H_0 .

Example 2.2.1 Consider the two-state Markov chain with its t.p.m. given by

$$\begin{matrix} & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix} \end{matrix}.$$

The null parameter space is $\Theta_0 = (0, 1)$ and for all θ , $D = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. The vector of derivatives of the elements of the t.p.m., taken row-wise is given by $(-1, 1, 1, -1)$, whose rank is 1. Besides, the chain is irreducible and aperiodic with $(1/2, 1/2)$ as the unique stationary distribution. It is straightforward that the MLE is given by $\hat{\theta} = (N_{01} + N_{10})/T$. The asymptotic null distribution of each of the statistics X^2 and LRT is χ_1^2 .

Example 2.2.2 Let $\Theta = (0, 1)$ and let the t.p.m. of a Markov chain be given by

$$\begin{pmatrix} 1 & 0 \\ \theta & 1-\theta \end{pmatrix}.$$

The Markov chain is reducible and not ergodic. One can directly study the behavior of the MLE and verify that it is not consistent, for any initial distribution.

Example 2.2.3 The hypothesis that $\mathfrak{X}$ is a sequence of i.i.d. random variables within the hypotheses that $\mathfrak{X}$ is a first order Markov chain can be represented as a parametric model where $p_{ij} = \pi_j$ for all i, j , where $\sum_j \pi_j = 1$. Obviously, $\hat{\pi}_j = N_j/T$. The LRT and the Pearson's X^2 statistics are respectively given by

$$-2 \ln \Lambda = -2 \sum_{i,j} \ln \left(\frac{T N_{ij}}{N_i N_j} \right)$$

$$X^2 = \sum_{i,j} \frac{(N_{ij} - N_i N_j / T)^2}{N_i N_j / T}.$$

Under H_0 , both have a chi-square distribution with $M(M-1) - M - 1 = (M-1)^2$ d.f. Similarity with testing for independence in a contingency table is obvious.

Example 2.2.4 Let us suppose that $\mathfrak{X}$ is a second order Markov chain. (For verification of various conditions, it may be noted that a second order Markov chain can be represented as a first order vector-valued Markov chain $\{Y_t, t \geq 0\}$ where $Y_t = (X_t, X_{t+1})$.) Let us assume that $p_{ijk} > 0$ for all i, j, k . The hypothesis that $\{X_t, t \geq 0\}$ is a first order Markov chain represents a parametric model with $p_{ijk} = p_{jk}$ for all i, j, k . This justifies the large sample distributions of the two statistics for testing the first order against (*within*) the second order, stated earlier.

Example 2.2.5 Consider the Markov chain with the t.p.m.

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{pmatrix} 0 & 3/4 & 0 & 1/4 & 0 \\ 0 & \theta & 1-\theta & 0 & 0 \\ 0 & \theta/2 & 1-\theta/2 & 0 & 0 \\ 0 & 0 & 0 & 1-\theta^2 & \theta^2 \\ 0 & 0 & 0 & \theta & 1-\theta \end{pmatrix} \end{matrix},$$

where $0 < \theta < 1$. It is easy to see that there are two persistent and minimal closed classes viz., $\{2, 3\}$ and $\{4, 5\}$. The state 1 is transient. The regularity conditions in terms of continuity and differentiability are satisfied. But the probabilistic conditions of a parametric model are not satisfied. It can be shown that the MLE is consistent, however its asymptotic distribution is a mixture of two normal distributions, if $X_0=1$.

Example 2.2.6 Testing for a specified stationary distribution

The hypothesis of interest is $H_0 : \pi_i = \pi_i(0)$ where $\pi_i(0)$, $i = 1, 2, \dots, M$ is the specified stationary distribution of the Markov chain. We construct the LRT for this problem.

Consider the simple case $M = 2$ first. Let $(\pi_0(0), \pi_1(0))$, with $\pi_0(0) + \pi_1(0) = 1$, be the stationary distribution under H_0 . For the sake of notational convenience, let us denote the specified stationary distribution by (π_0, π_1) . The log-likelihood

$$N_{00} \ln p_{00} + N_{01} \ln p_{01} + N_{10} \ln p_{10} + N_{11} \ln p_{11}$$

needs to be maximized subject to the constraints that

$$(1) \pi_0 p_{00} + (1 - \pi_0) p_{10} = \pi_0$$

$$(2) \pi_0(1 - p_{01}) + (1 - \pi_0)(1 - p_{11}) = 1 - \pi_0.$$

However, since $p_{00} + p_{01} = 1$ and $p_{10} + p_{11} = 1$, we need to consider only one of these two constraints. We take the constraint (1). By the Lagrange's multiplier theorem, we set

$$g = N_{00} \ln p_{00} + N_{01} \ln(1 - p_{00}) + N_{10} \ln p_{10} + N_{11} \ln(1 - p_{10}) \\ + \lambda(\pi_0 p_{00} + (1 - \pi_0)p_{10} - \pi_0).$$

Then, we have the following equations to get ML estimator under H_0

$$\begin{aligned} \frac{\partial g}{\partial p_{00}} &= \frac{N_{00}}{p_{00}} - \frac{N_{01}}{1 - p_{00}} + \lambda \pi_0 = 0. \\ \frac{\partial g}{\partial p_{10}} &= \frac{N_{10}}{p_{10}} - \frac{N_{11}}{1 - p_{10}} + (1 - \pi_0)\lambda = 0. \\ \frac{\partial g}{\partial \lambda} &= \pi_0 p_{00} + (1 - \pi_0)p_{10} - \pi_0 = 0. \end{aligned}$$

The above system can be solved by an iterative scheme. For an M state Markov chain, we set

$$g = \sum_{i,j=1}^M N_{ij} \ln p_{ij} + \sum_{i=1}^{M-1} \lambda_i \left(\pi_{0i} - \sum_{j=1}^M \pi_{0j} p_{ji} \right) + \sum_{i=1}^M \eta_i \left(\sum_{j=1}^M p_{ij} - 1 \right),$$

where λ_i 's and η_j 's are the Lagrangian parameters. The LRT has a large sample χ_{M-1}^2 distribution, under H_0 .

Another strategy is to construct the Wald's test. We compute the stationary distribution of $\hat{P}$, the MLE of P and compare it with $\pi_i(0)$. Let $\hat{\pi}_i$, $i = 1, 2, \dots, M$ be the stationary distribution of $\hat{P}$. To construct the Wald's test for H_0 , we need to compute the variance-covariance matrix of $\hat{\pi}_i$, $i = 1, 2, \dots, M$.

There are three vectors to be compared: the observed proportions of various states, $\hat{\pi}_i$, $i = 1, 2, \dots, M$ (the stationary distribution of $\hat{P}$) and $\pi_i(0)$. The first vector is almost the same as the stationary distribution of $\hat{P}$ for a large T .

It is more convenient to construct a Wald's test based on a quadratic form in observed proportions of various states. Let $U_t(i) = I[X_t = i]$, $i = 1, 2, \dots, M$, $t = 1, 2, \dots, T$. Then, $\bar{\pi}_i = \bar{U}_T(i) = \sum_{t=1}^T I[X_t = i]/T$. Let $p_{ij}^{(t)} = P[X_t = j | X_0 = i]$ denote a t -step transition probability. Then,

$$\begin{aligned} \text{Cov}(\bar{U}_T(i), \bar{U}_T(j)) &= \frac{1}{T^2} \sum_{s=1}^T \sum_{t=1}^T \text{Cov}(U_s(i), U_t(j)) \\ &= \frac{1}{T^2} \sum_{s=1}^T \sum_{t=1}^T [\pi_i p_{ij}^{(|t-s|)} - \pi_i \pi_j]. \end{aligned}$$

The stationary probabilities can be estimated by the observed proportions $\tilde{\pi}_i$'s. Powers of $\hat{P}$ can be used to estimate $p_{ij}^{(t)}$'s in the above.

Another and possibly simpler procedure to estimate the above variances and covariances is as follows. In view of the stationarity, the covariance $\text{Cov}(U_s(i), U_t(j))$ is a function of $|t - s|$ only. Consider then $\text{Cov}(U_1(i), U_d(j))$, $d \geq 1$ which can be estimated by

$$\hat{\text{Cov}}(U_1(i), U_d(j)) = \frac{1}{T-d-L} \sum_{s=1}^{T-d-L} (U_s(i) - \bar{U}(1, i)) (U_{s+d+L}(j) - \bar{U}(2, j)), \quad (2.1)$$

where $\bar{U}(1, i) = \frac{1}{T-d-L} \sum_{s=1}^{T-d-L} U_s(i)$ and $\bar{U}(2, j) = \frac{1}{T-d-L} \sum_{s=1}^{T-d-L} U_{s+d+L}(j)$ and $L = L(T)$ is a sequence of integers such that $L \rightarrow \infty$ and $\sqrt{L}/T \rightarrow 0$. In practice, for large T , the two means $\bar{U}(1, i)$ and $\bar{U}(2, j)$ can be replaced by the mean of all the observations, however, in such a case, it is possible that for some samples, the corresponding estimator of the variance-covariance matrix of $(\bar{U}_T(1), \bar{U}_T(2), \dots, \bar{U}_T(M))$ is not non-negative definite. We notice that the above estimator does not depend on the assumed Markov model and it can be shown to be consistent for sequences more general than Markov chains. Under the assumption that the covariances of larger lags are negligible, it focuses on the more dominant terms of the covariance. It is a special case of estimators that we discuss in some more detail in Sect. 6.6. Under the assumptions that we have made, the Markov chain is geometrically ergodic (i.e., $|p_{ij}^{(t)} - \pi_j| < C\rho^t$, $0 \leq \rho < 1$). As remarked in Example 1.3.2, it is Geometrically Strong Mixing (It can be shown that such an estimator of the variance is consistent, cf. Theorem 6.6.1).

Let $\hat{\Sigma}$ be a consistent estimator of the variance covariance of the vector of observed proportions of states $\tilde{\Pi} = (\tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_M)'$ and let $\Pi(0) = (\pi_1(0), \pi_2(0), \dots, \pi_M(0))'$. Then, the Wald test-statistic is given by

$$X^2 = T \left(\tilde{\Pi} - \Pi(0) \right)' \hat{\Sigma}^+ \left(\tilde{\Pi} - \Pi(0) \right), \quad (2.2)$$

where A^+ denotes the Moore-Penrose g-inverse of a matrix A . Under H_0 , the test-statistic X^2 has asymptotically a χ_{M-1}^2 distribution under H_0 .

Markov chains with infinitely many states

We take the state-space as $S = \{0, 1, 2, \dots\}$. Under the assumptions of Theorem 1.1 of Billingsley (1961), it follows that there is a consistent solution of the likelihood equations which is asymptotically normal with mean vector 0 and the variance-covariance matrix $[I(\theta)]^{-1}$.

Example 2.2.7 Poisson Markov Sequence

A Poisson-Markov sequence $\{X_t, t = 0, 1, \dots\}$ is defined as follows:

- A1. $\{Y_t, t = 0, 1, \dots\}$ is a sequence of i.i.d. Poisson random variables with $E(Y_0) = \lambda$.
- A2. $P[Z_{t+1} = z | X_t = x, \dots] = \binom{x}{z} p^z (1-p)^{x-z}$
- A3. Y_t is independent of $X_0, X_1, \dots, X_{t-1}, Z_0, Z_1, \dots, Z_t$, for each t .
- A4. $X_{t+1} = Z_{t+1} + Y_{t+1}$.

In applications of a Poisson Markov sequence, X_t stands for the total number of members of the population at time t and Y_t are new recruits or newly joining members, whereas the random variable Z_t denotes survivors from the earlier day. Assumption A2 is equivalent to the assumption that an existing member survives for yet another time unit with probability p , irrespective of its age and independently of other members of the population. Assumption A3 says that Y_t , the number of new arrivals, is independent of the existing and past members (density-independent recruitment). This is, in fact, a discrete version of the $M|M|\infty$ queuing system. It follows that the one-step transition probability is given by

$$p_{ij} = P[X_{t+1} = j | X_t = i] = \sum_{z=0}^{\min(i,j)} \binom{i}{z} p^z (1-p)^{i-z} \frac{e^{-\lambda} \lambda^{j-z}}{(j-z)!}$$

We notice that $p_{ij} > 0 \forall (i, j)$, thus the Markov chain is irreducible and aperiodic. Further, p_{ij} is thrice differentiable in p and λ .

Now,

$$E(X_{t+1} | X_t) = E(Z_{t+1} + Y_{t+1} | X_t) = pX_t + \lambda.$$

If we assume that the process is stationary, $E(X_{t+1}) = E(X_t) = \mu$ (say), which implies that $\mu p + \lambda = \mu$ or $\mu = \lambda / (1 - p)$. A similar argument for variance implies that $\text{Var}(X_t)$ also equals μ for all t , if we assume stationarity. This suggests that the stationary distribution of the process is Poisson with mean μ . This is proved based on the following result, which is easy to prove.

Lemma 2.2.1 *If U has a Poisson distribution with mean λ , and if the distribution of V given $U = u$ is Binomial(u, p) then V is Poisson with mean λp (if $U = 0$, we define $V = 0$).*

We recall the following well-known result for Markov chains.

Theorem 2.2.1 *A Markov chain $\{X_t, t = 0, 1, \dots\}$ is strictly stationary, if and only if, X_0 and X_1 are identically distributed. Their common distribution is a stationary distribution.*

Proof Let $P[X_0 = j] = p_j, j \in S$. Then, $P[X_1 = j] = \sum_i P[X_0 = i, X_1 = j] = \sum_i p_i p_{ij}$. That is, $p_j = \sum_i p_i p_{ij} \forall j$, which satisfies the Definition 1.1.4 of a stationary distribution.

Now suppose that $X_0 \sim \text{Poisson}(\eta)$. Therefore, $X_1 \sim \text{Poisson}(\eta p + \lambda)$ by the Lemma and the Assumptions A1 and A2. Then X_0 and X_1 are identically distributed

if and only if $\eta = \eta p + \lambda$. Thus, if $\eta = \lambda/(1 - p)$, the stationary distribution of the process is Poisson with mean $\mu = \lambda/(1 - p)$. Hence, the process is non-null persistent for all λ, p . We thus see that all the assumptions of a parametric model are satisfied.

Case I. Suppose $\{X_t, Y_t\}$, $t = 0, 1, \dots, T$ are both observed. In this case, maximum likelihood estimation of parameters is very easy to carry out.

Case II. Now suppose that only X_t 's have been observed. We re-parametrize the model in terms of μ and p . In this case, one can show that $\bar{X}$ is a good approximation to the MLE of μ (see Guttorp (1995), page 100). Estimation of p needs to be carried out by using iterative numerical procedures, such as Newton-Raphson. The Fisher Information matrix is rather involved and we may use the matrix $\hat{F}$ as its estimator.

2.3 Extensions of Markov Chain Models

Models based on the Logistic Regression.

Consider a two-state Markov chain. Let us write

$$\ln \frac{P[X_{t+1} = 1|X_t = 0]}{P[X_{t+1} = 0|X_t = 0]} = \beta_0 \text{ (base-line probability)}$$

$$\text{and} \quad \ln \frac{P[X_{t+1} = 1|X_t = 1]}{P[X_{t+1} = 0|X_t = 1]} = \beta_0 + \beta_1.$$

The t.p.m. can be written as

$$P = \begin{pmatrix} \frac{1}{1+e^{\beta_0}} & \frac{e^{\beta_0}}{1+e^{\beta_0}} \\ \frac{1}{1+e^{\beta_0+\beta_1}} & \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \end{pmatrix}.$$

Though this appears to be only a re-parametrization, it serves to be useful and convenient. The above model in a compact form is written as

$$\ln \frac{P[X_{t+1} = 1|X_t]}{P[X_{t+1} = 0|X_t]} = \beta_0 + \beta_1 X_t.$$

The second order Markov chain with two states is modeled as

$$\ln \frac{P[X_{t+1} = 1|X_t, X_{t-1}]}{P[X_{t+1} = 0|X_t, X_{t-1}]} = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1}.$$

The number of parameters in the above model is 3, whereas the saturated second order two-state Markov chain has 4 parameters. However, it must be pointed out that in such a model, unlike the saturated model, the transition probabilities are functions

of the coding or numerical labels used to denote states of the chain. If the state-space refers to say linguistic classes, such as consonants and vowels, the above model need not be realistic.

An L order Markov Chain can be similarly defined by

$$\ln \frac{P[X_t = 1|X_{t-1}, X_{t-2}, \dots, X_{t-L}]}{P[X_t = 0|X_{t-1}, X_{t-2}, \dots, X_{t-L}]} = \beta_0 + \sum_{l=1}^L \beta_l X_{t-l}.$$

The above model has $L+1$ parameters as opposed to the saturated model which has 2^L parameters. A further advantage of such an approach is that we can incorporate time-dependent regressors also. Let $\{z_t\}$ be the sequence of values of regressors (possibly vector valued). Either the sequence $\{z_t\}$ is deterministic or if it is stochastic, the model is a conditional one. We write

$$\ln \frac{P[X_t = 1|X_{t-1}, X_{t-2}, \dots, X_{t-L}, z_t]}{P[X_t = 0|X_{t-1}, X_{t-2}, \dots, X_{t-L}, z_t]} = \beta_0 + \sum_{l=1}^L \beta_l X_{t-l} + \gamma' z_t.$$

A major advantage of this approach is that we can use any statistical package which analyzes logistic regression models.

M state L order Markov chain it based on Polytomous regression model. The logistic regression model for two categories can be extended to several categories, which is known as polytomous or multinomial (logistic) regression model. A polytomous regression model can be used to define an L order Markov chain with M states. Let $\tilde{P}[X_{L+1} = i] = P[X_{L+1} = i|X_1, X_2, \dots, X_L]$, $i = 1, 2, \dots, M$. Then,

$$\ln \frac{\tilde{P}[X_{L+1} = i]}{\tilde{P}[X_{L+1} = M]} = \beta_{0i} + \beta_{1i} X_1 + \dots + \beta_{Li} X_L, \quad i = 1, 2, \dots, M-1.$$

$$\tilde{P}[X_{L+1} = M] = \frac{1}{1 + \sum_{i=1}^{M-1} \sum_{\ell=1}^L \exp\{\beta_{0i} + \beta_{\ell i} X_\ell\}}.$$

The above model has $(L+1)(M-1)$ parameters, far less than the corresponding saturated L -order model, which has $M^L(M-1)$ parameters. Thus, such a model has the advantage of having less parameters and any software which analyzes the polytomous logistic regression data can be easily used to get the maximum likelihood estimators and estimators of their asymptotic variance-covariance matrix. Most of the packages include tests for significance of regression parameters. Such procedures can be used for testing of an order of a Markov chain. Analysis of higher order Markov chain models can also be carried out based on log-linear contingency table models, see Davison (2003).

Raftery's Mixture Transition Distribution Model

As observed earlier, a higher order M -state Markov chain model has too many parameters. An important higher order model with a significantly less number of

parameters is due to Raftery (1985) and it is known as the Mixture Transition Distribution (MTD) model. The MTD model of order L is defined by

$$P[X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_{t-L} = x_{t-L}, \dots, X_1 = x_1, X_0 = x_0] \\ = \sum_{l=1}^L \lambda_l p_{x_{t-l}x_t},$$

whenever the conditional probability on the left-hand side is defined. In the above, $P = ((p_{ij}))$ is an $M \times M$ stochastic matrix and $\{\lambda_l, l = 1, 2, \dots, L\}$ constitutes a probability distribution. A probabilistic interpretation of the above model is as follows. Nature chooses the lag l with probability λ_l . If the chosen lag is r and if $X_{t-r} = i$, the conditional probability of $X_t = j$ given the chosen lag and the entire past is p_{ij} . The number of parameters of an MTD model of order L is $(L - 1) + M(M - 1)$, far less than the corresponding saturated model which has $M^L(M - 1)$ parameters. Adke and Deshmukh (1988) have shown that, if the matrix P is positively regular, i.e., if there exists a t such that all the elements of P^t are positive, the MTD model has the property that $P[X_{t+s} = j | X_s = i]$ converges to π_j , as $t \rightarrow \infty$ for every i, j , where π_j is, in fact, the unique stationary distribution of a Markov chain whose one-step t.p.m. is given by P . If X_0 follows the distribution π_j , the MTD model is stationary with π_j as the common distribution of X_t for all t . This result is useful in establishing properties of the MLE. Further, the MTD model is a parametric Markov model of order L and it can be shown to satisfy the Cramer regularity conditions. For a detailed discussion of MTD models including numerical procedures for estimation of parameters, we refer to Berchtold and Raftery (2002).

2.4 Hidden Markov Chains

Let $\{Y_t, t \geq 1\}$ be a stationary, irreducible, and aperiodic Markov chain on the state-space $\{1, 2, \dots, M\}$ with P as the one-step t.p.m. and $\{\pi_i, i = 1, 2, \dots, M\}$ as the unique stationary distribution. The chain $\{Y_t, t \geq 1\}$ is not observable. We observe the process $\{X_t, t \geq 1\}$, the state-space of which is $\{1, 2, \dots, N\}$. The conditional distribution of X_t is given by

$$P[X_t = k | Y_t = j, Y_{t-1}, \dots, Y_1, X_{t-1}, X_{t-2}, \dots, X_1] = P[X_t = k | Y_t = j] = q_{jk},$$

where $Q = ((q_{jk}))$ is an $M \times N$ matrix. The random sequence $\{X_t, t \geq 1\}$ on the state-space $\{1, 2, \dots, N\}$ is known as a Hidden Markov chain. (In literature, $\{(X_t, Y_t) t \geq 1\}$ is sometimes described as a Hidden Markov chain.) In general, $\{X_t, t \geq 1\}$ does not satisfy Markov property. There are three important issues to be addressed.

- (1) To derive likelihood function, i.e., $P[X_1 = x_1, X_2 = x_2, \dots, X_T = x_T]$.

- (2) To carry out state estimation, i.e., to derive the predictive distribution of the underlying chain :

$$P[Y_t = y_t | X_1 = x_1, \dots, X_T = x_T], \quad t = 1, 2, \dots, T.$$

Prediction of Y_{T+j} , $j \geq 1$ may also be of interest.

- (3) To carry out maximum likelihood estimation of the two unknown matrices P and Q

It is easy to see why (1) is involved: the likelihood function is a sum over T variables which correspond to the unobserved states of Y_t , $t = 1, 2, \dots, T$. However, there exist recursive algorithms for (2) and (3) above, which are easy to implement.

Forward Algorithm. Define

$$\begin{aligned} \alpha_i(t) &= P[X_1 = x_1, \dots, X_{t-1} = x_{t-1}, Y_t = i], \quad t = 2, 3, \dots, T, \quad i = 1, 2, \dots, M, \\ \alpha_i(1) &= P[Y_1 = i] = \pi_i. \end{aligned}$$

The last equation defines the initialization of the algorithm. If any of the event $X_1 = x_1, \dots, X_{t-1} = x_{t-1}$ is not well-defined, we take that event as Ω instead of ϕ , the empty set. In a recursive algorithm, we assume that $\alpha_i(1)$ are given for all i and find $\alpha_i(t)$ for $\forall i$ and $\forall t = 2, \dots, T$. We further define

$$\alpha_i(T+1) = P[X_1 = x_1, X_2 = x_2, \dots, X_T = x_T, Y_{T+1} = i].$$

Then, it is easily seen that the likelihood function is given by

$$P[X_1 = x_1, X_2 = x_2, \dots, X_T = x_T] = \sum_{i=1}^M \alpha_i(T+1). \quad (2.3)$$

Next, by the defining properties of the two processes, we get

$$\begin{aligned} \alpha_i(2) &= P[X_1 = x_1, Y_2 = i] \\ &= \sum_{j=1}^M P[X_1 = x_1, Y_1 = j, Y_2 = i] \\ &= \sum_{j=1}^M \alpha_j(1) q_{jx_1} p_{ji}. \end{aligned}$$

In general,

$$\alpha_i(t+1) = \sum_{j=1}^M \alpha_j(t) q_{jx_t} p_{ji}. \quad (2.4)$$

Next, we discuss the following Backward algorithm.

Backward Algorithm. Let

$$\beta_i(t) = P[X_t = x_t, \dots, X_T = x_T | Y_t = i], \quad t = 1, 2, \dots, T, \quad i = 1, 2, \dots, M,$$

$$\beta_i(T+1) = 1.$$

We note that

$$\begin{aligned} \beta_i(1) &= P[X_1 = x_1, \dots, X_T = x_T | Y_1 = i] \\ \beta_i(1)\pi_i &= P[X_1 = x_1, \dots, X_T = x_T, Y_1 = i], \\ \beta_i(T) &= P[X_T = x_T | Y_T = i] = q_{ix_T}. \end{aligned}$$

The Backward algorithm involves expressing $\beta_i(t)$ in term of $(\beta_i(t+1), \dots, \beta_M(t+1))$. We observe that

$$\begin{aligned} \beta_i(t) &= P[X_t = x_t, \dots, X_T = x_T | Y_t = i] \\ &= P[X_t = x_t, \dots, X_T = x_T, Y_t = i] / \pi_i. \end{aligned}$$

Then, from the properties of the Hidden Markov chain,

$$\begin{aligned} \beta_i(t) &= \sum_{j=1}^M P[X_t = x_t, X_{t+1} = x_{t+1}, \dots, X_T = x_T, Y_{t+1} = j | Y_t = i] \\ &= \sum_{j=1}^M P[X_t = x_t, X_{t+1} = x_{t+1}, \dots, X_T = x_T, Y_t = i, Y_{t+1} = j] / P[Y_t = i] \\ &= \sum_{j=1}^M P[Y_t = i, X_t = x_t, Y_{t+1} = j, X_{t+1} = x_{t+1}, \\ &\quad X_{t+2} = x_{t+2}, \dots, X_T = x_T] / P[Y_t = i] \\ &= \sum_{j=1}^M P[Y_t = i] P[X_t = x_t | Y_t = i] P[Y_{t+1} = j | X_t = x_t, Y_t = i] \\ &\quad P[X_{t+1} = x_{t+1}, \dots, X_T = x_T | Y_{t+1} = j, X_t = x_t, Y_t = i] / P[Y_t = i] \\ &= \sum_{j=1}^M q_{ix_t} p_{ij} \beta_{t+1}(j). \end{aligned}$$

Let $\mathbf{X} = (X_1, X_2, \dots, X_T)$ and $\mathbf{x} = (x_1, x_2, \dots, x_T)$. Combining Backward and Forward Algorithms, we have

$$\begin{aligned}
 & P[X_1 = x_1, \dots, X_t = x_t, X_{t+1} = x_{t+1}, \dots, X_T = x_T, Y_t = i] \\
 &= P[X_1 = x_1, \dots, Y_t = i, X_t = x_t, X_{t+1} = x_{t+1}, \dots, X_T = x_T] \\
 &= P[X_1 = x_1, \dots, X_{t-1} = x_{t-1}, Y_t = i] \\
 &\quad \times P[X_t = x_t, X_{t+1} = x_{t+1}, \dots, X_T = x_T | Y_t = i] \\
 &= \alpha_i(t) \beta_i(t).
 \end{aligned}$$

Thus,

$$P[\mathbf{X} = \mathbf{x}, Y_t = i] = \alpha_i(t) \beta_i(t)$$

and therefore,

$$P[\mathbf{X} = \mathbf{x}] = \sum_{i=1}^M \alpha_i(t) \beta_i(t).$$

Now, if we had observed $Y_1, \dots, Y_T$ also, the ML estimates of p_{ij} and q_{ik} would have been $\hat{p}_{ij} = \frac{\sum_t I[Y_t=i, Y_{t+1}=j]}{\sum_t I[Y_t=i]}$ (cf. Sect. 1.1) and $\hat{q}_{ik} = \frac{\sum_t I[Y_t=i, X_t=k]}{\sum_t I[Y_t=i]}$ respectively. The Baum-Welsh algorithm which we discuss below, computes conditional expectation of $I[Y_t = i, Y_{t+1} = j]$ and $I[Y_t = i, X_t = k]$ given the observations $\mathbf{X}$. We have

$$\begin{aligned}
 p_t(i, j) &= P[Y_t = i, Y_{t+1} = j | \mathbf{X} = \mathbf{x}] \\
 &= \frac{P[Y_t = i, Y_{t+1} = j, \mathbf{X} = \mathbf{x}]}{P[\mathbf{X} = \mathbf{x}]}.
 \end{aligned}$$

Now, from the properties of a Hidden Markov chain,

$$\begin{aligned}
 & P[Y_t = i, Y_{t+1} = j, \mathbf{X} = \mathbf{x}] \\
 &= P[X_1 = x_1, \dots, X_{t-1} = x_{t-1}, Y_t = i, X_t = x_t, Y_{t+1} = j, \\
 &\quad X_{t+1} = x_{t+1}, \dots, X_T = x_T] \\
 &= \alpha_i(t) q_{ix_t} p_{ij} \beta_j(t+1).
 \end{aligned}$$

Hence, the likelihood is given by

$$P[\mathbf{X} = \mathbf{x}] = \sum_{\ell=1}^M \sum_{k=1}^M \alpha_\ell(t) q_{\ell x_t} p_{\ell k} \beta_k(t+1).$$

and

$$p_t(i, j) = \frac{\alpha_i(t) q_{ix_t} p_{ij} \beta_j(t+1)}{\sum_{\ell=1}^M \sum_{k=1}^M \alpha_\ell(t) q_{\ell x_t} p_{\ell k} \beta_k(t+1)}. \quad (2.5)$$

We may observe that the likelihood is also given by

$$L(P, Q) = P[X_1 = x_1, \dots, X_T = x_T] = \sum_{i=1}^M \beta_i(1) \pi_i. \quad (2.6)$$

We do not use the likelihood (2.3) or (2.6) for ML estimation, however, it is needed while comparing the Hidden Markov model with competing models. The recursive algorithm is as follows. The current estimate of p_{ij} is given by

$$\hat{p}_{ij} = \frac{\sum_t p_t(i, j)}{\sum_j \sum_t p_t(i, j)}.$$

Let

$$\gamma_i(t) = \sum_{j=1}^M p_t(i, j)$$

be the probability that the state i is observed at time t . Let $A(k) = \{t | I[X_t = k]\}$. The current estimate of q_{ik} is then given by

$$\hat{q}_{ik} = \frac{\sum_{t \in A(k)}^N \gamma_i(t)}{\sum_{t=1}^N \gamma_i(t)}. \quad (2.7)$$

We begin the algorithm with arbitrary estimates of the matrices P and Q and update their values as given above. The procedure continues until successive values differ by a pre-assigned tolerance. The algorithm is known as Baum-Welch or Forward-Backward algorithm. It is, in fact, one of the early versions of the EM algorithm, frequently used in the context of incomplete observations or samples with missing data.

For State Estimation, we need to obtain

$$\begin{aligned} & \arg \max_{y_1, \dots, y_T} P[Y_1 = y_1, \dots, Y_T = y_T | X_1 = x_1, \dots, X_T = x_T]. \\ &= \arg \max_{y_1, \dots, y_T} P[Y_1 = y_1, \dots, Y_T = y_T, X_1 = x_1, \dots, X_T = x_T]. \end{aligned}$$

To do so, we apply the *Viterbi Algorithm* which is also recursive in nature. Let

$$\delta_j(t) = \max_{y_1, y_2, \dots, y_{t-1}} P[Y_1 = y_1, \dots, Y_{t-1} = y_{t-1}, Y_t = j, X_1 = x_1, \dots, X_{t-1} = x_{t-1}].$$

The initialization is carried out by

$$\delta_j(1) = \pi_j, \quad j = 1, 2, \dots, M,$$

the stationary distribution of the chain $\{Y_t, t \geq 1\}$. It can be shown that

$$\delta_j(t+1) = \max_{i=1,2,\dots,M} [\delta_i(t) p_{ij} q_{jx_t}].$$

Now, let

$$\psi_j(t+1) = \arg \max_{i=1,2,\dots,M} [\delta_i(t) p_{ij} q_{jx_t}].$$

The variable $\psi_j(t)$ records the “node of the incoming arc” that has resulted in this most probable path. The algorithm terminates with

$$\hat{Y}_{T+1} = \max_{i=1,2,\dots,M} \delta_i(T+1)$$

and

$$\hat{Y}_t = \psi_{\hat{Y}_{t+1}}(t+1),$$

which are predictors of $Y_1, Y_2, \dots, Y_T$. In the above algorithm, if there are ties, we break them randomly. Also, the algorithm assumes that the model parameters are known. In practice, it is implemented by replacing the unknown parameters (P, Q) by their MLEs.

The above discussion of the Backward-Forward algorithm is based on Manning and Schütze (1999). We refer to MacDonald and Zucchini (1997) for a thorough account of Hidden Markov processes on general state-spaces and their applications.

2.5 Aggregate Data from Finite Markov Chains

Model I. Here, we observe N i.i.d. finite Markov chains, each having the t.p.m. P . At each time unit $t = 1, 2, \dots, T$, we observe the number of Markov chains in the state i . However, transitions from a state i to a state j of these individual Markov chains are not available. Let $N(t, i)$ = Number of units or Markov chains in the state i at time t .

Notice that $\sum_i N(t, i) = N \quad \forall t$. It is easy to see that

$$E[N(t, j) | N(t-1, 1), \dots, N(t-1, M)] = \sum_{i=1}^M N(t-1, i) p_{ij}.$$

This leads to a regression model, where the vector Y is the responses $N(t, j)$'s for $t = 2, 3, \dots, N$ and $j = 1, 2, \dots, M$. The regression vector β is the transition probabilities p_{ij} 's written in a conveniently chosen column form. Random variables

$N(t-1, i)$'s act as regressors to lead to a regression setup $E[Y] = X\beta$, in standard notations of regression analysis. We then have the Ordinary Least Square (OLS) estimator $\hat{\beta} = (X'X)^{-1}X'Y$. (One may take only the first $M-1$ elements of each row of P and $N(t, i)$ for $i = 1, 2, \dots, M-1$). We note that variances of the response variables are not the same, also, they are not independent. We then need to consider the Weighted Least Squares (WLS). Both OLS and WLS estimators ignore the property that the transition probabilities are non-negative. (It is known that the OLS satisfies the condition that each of the row sum is 1, (cf. Lee et al. (1970), p. 34)) Thus, a better strategy is to minimize

$$(Y - X\beta)'(Y - X\beta)$$

subject to the constraints (i) $p_{ij} \geq 0, \forall (i, j)$ (ii) $\sum_{j=1}^M p_{ij} = 1 \forall i$. This is a constrained optimization problem (in fact, a Quadratic Programming Problem) and it can be shown to have a unique solution. A software package such as MATLAB or GAUSS can be used to get a solution to the optimization problem.

It can be shown that the sequence of $M \times 1$ random vectors $\{(N(t, 1), \dots, N(t, M)), t \geq 1\}$ forms a Markov chain. Further, the conditional distribution of the random vector $(N(t, 1), N(t, 2), \dots, N(t, M))$ is a multinomial distribution with parameters N and the cell probabilities $\sum_i (N(t-1, i)/N) p_{i1}, \sum_i (N(t-1, i)/N) p_{i2}, \dots, \sum_i (N(t-1, i)/N) p_{iM}$. Verification of the regularity conditions is straightforward. *Model II.* It is not necessary that we observe the same N individuals throughout. Thus, at each time, we observe $N(t, i), i = 1, 2, \dots, M$ and $\sum_i N(t, i) = N(t)$ which need not be the same for all t . Technically, the T random vectors

$$\{N(1, 1), \dots, N(1, M)\}, \{N(2, 1), \dots, N(2, M)\}, \dots, \{N(T, 1), \dots, N(T, M)\}$$

are independently distributed. Now,

$$E[N(t+1, j)] = \sum_i \pi_{t,i} p_{ij},$$

where $\pi_{t,i}$ is the probability that a randomly chosen person (at time t) is in the state i . This is a case of moment estimation, where we first estimate $\pi_{t,i}$ by the observed proportions $N(t, i)/N(t)$. Replacing this estimate in the above, we again get the situation similar to Model I and employ the LSEs. Lee et al. (1970) have an extensive review of the various methods to estimate the transition probabilities.

In each of the above cases, we can allow either T to tend to ∞ or $N(t)$ to tend to ∞ for each t (or both). In either case, it can be shown that LSE/MLE is consistent and asymptotically normal with appropriate norming. When the process reaches equilibrium, for large N , the relative frequencies of M states at time t are close to the unique stationary distribution and therefore to each other. Since they act as regressors, the matrix X of the above regression model turns out to be nearly singular. This leads to a multi-collinearity problem. Inderdeep Kaur and Rajarshi (2012) discuss

ridge-regression estimators which offer a considerable improvement over the LSE in terms of the total mean squared error.

Books by Guttorp (1995), Davison (2003) (Chap. 6) and Lindsey (2004) are good sources of inference in Markov chains and related topics.

References

- Adke, S.R., Deshmukh, S.R.: Limit distribution of a higher order Markov chains. *J. Roy. Stat. Soc. Ser. B* **50**, 105–108 (1988)
- Berchtold, A., Raftery, A.E.: Mixture transition distribution model for high-order markov chains and non-gaussian time series. *Stat. Sci.* **17**, 328–356 (2002)
- Billingsley, P.: *Statistical Inference for Markov Processes*. The University of Chicago Press, Chicago (1961)
- Cappé, O., Moulines, E., Ryden, T.: *Inference in Hidden Markov Models*. Springer, New York (2005)
- Davison, A.C.: *Statistical Models*. Cambridge University Press, Cambridge (2003)
- Elliot, R.J., Aggaoun, L., Moore, J.B.: *Hidden Markov Models: Estimation and Control*. Springer, New York (1995)
- Guttorp, P.: *Stochastic Modeling of Scientific Data*. Chapman and Hall, London (1995)
- Inderdeep Kaur, Rajarshi: Ridge regression for estimation of transition probabilities from aggregate data. *Commun. Stat. Simul. Comput.* **41**, 524–530 (2012)
- Katz, R.W.: On some criteria for estimating the order of a markov chain. *Technometrics* **23**, 243–249 (1981)
- Lee, T.C., Judge, G.G., Zellner, A.: *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*. North Holland, Amsterdam (1970)
- Lindsey, J.K.: *Statistical Analysis of Stochastic Processes in Time*. Cambridge University Press, Cambridge (2004)
- MacDonald, I., Zucchini, W.: *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London (1997)
- Manning, C.D., Schütze, H.: *Statistical Natural Language Processing*. MIT Press, Cambridge (1999)
- Raftery, A.E.: A model for high order Markov chains. *J. Roy. Stat. Soc. Ser. B* **47**, 528–539 (1985)



<http://www.springer.com/978-81-322-0762-7>

Statistical Inference for Discrete Time Stochastic
Processes

Rajarshi, M.B.

2013, XI, 113 p., Softcover

ISBN: 978-81-322-0762-7