6.2.2 Parameters Estimation

In this subsection, we present efficient methods to estimate the parameters Q_i and λ_i for $i = 1, 2, \ldots, k$. To estimate Q_i, one may regard Q_i as the i-step transition matrix of the categorical data sequence $\{X^{(n)}\}$. Given the categorical data sequence $\{X^{(n)}\}$, one can count the transition frequency $f_{jl}^{(i)}$ in the sequence from State l to State j in the i-step. Hence one can construct the i-step transition matrix for the sequence $\{X^{(n)}\}$ as follows:

$$F^{(i)} = \begin{pmatrix} f_{11}^{(i)} & \cdots & \cdots & f_{m1}^{(i)} \\ f_{12}^{(i)} & \cdots & \cdots & f_{m2}^{(i)} \\ \vdots & \vdots & \vdots & \vdots \\ f_{1m}^{(i)} & \cdots & \cdots & f_{mm}^{(i)} \end{pmatrix}. \tag{6.9}$$

From $F^{(i)}$, we get the estimates for $Q_i = [q_{lj}^{(i)}]$ as follows:

$$\hat{Q}_i = \begin{pmatrix} \hat{q}_{11}^{(i)} & \cdots & \cdots & \hat{q}_{m1}^{(i)} \\ \hat{q}_{12}^{(i)} & \cdots & \cdots & \hat{q}_{m2}^{(i)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{q}_{1m}^{(i)} & \cdots & \cdots & \hat{q}_{mm}^{(i)} \end{pmatrix} \tag{6.10}$$

where

$$\hat{q}_{lj}^{(i)} = \begin{cases} \dfrac{f_{lj}^{(i)}}{\displaystyle\sum_{l=1}^{m} f_{lj}^{(i)}} & \text{if } \displaystyle\sum_{l=1}^{m} f_{lj}^{(i)} \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6.11}$$

We note that the computational complexity of the construction of $F^{(i)}$ is of $O(L^2)$ operations, where L is the length of the given data sequence. Hence the total computational complexity of the construction of $\{F^{(i)}\}_{i=1}^{k}$ is of $O(kL^2)$ operations. Here k is the number of lags.

The following proposition shows that these estimators are unbiased.

Proposition 6.2. *The estimators in (6.11) satisfies*

$$E(f_{lj}^{(i)}) = q_{lj}^{(i)} E\left(\sum_{j=1}^{m} f_{lj}^{(i)}\right).$$

Proof. Let T be the length of the sequence, $[q_{lj}^{(i)}]$ be the i-step transition probability matrix and $\bar{X}_l$ be the steady state probability that the process is in state l. Then we have

$$E(f_{lj}^{(i)}) = T \cdot \bar{X}_l \cdot q_{lj}^{(i)}$$

and

$$E(\sum_{j=1}^{m} f_{lj}^{(i)}) = T \cdot \bar{X}_l \cdot (\sum_{j=1}^{m} q_{lj}^{(i)}) = T \cdot \bar{X}_l.$$

Therefore we have

$$E(f_{lj}^{(i)}) = q_{lj}^{(i)} \cdot E(\sum_{j=1}^{m} f_{lj}^{(i)}).$$

In some situations, if the sequence is too short then $\hat{Q}_i$ (especially $\hat{Q}_k$) contains a lot of zeros (therefore $\hat{Q}_n$ may not be irreducible). However, this did not occur in the tested examples. Here we propose the second method for the parameter estimation. Let $\mathbf{W}^{(i)}$ be the probability distribution of the i-step transition sequence, then another possible estimation for Q_i can be $\mathbf{W}^{(i)}\mathbf{1}^t$. We note that if $\mathbf{W}^{(i)}$ is a positive vector, then $\mathbf{W}^{(i)}\mathbf{1}^t$ will be a positive matrix and hence an irreducible matrix.

Proposition 6.1 gives a sufficient condition for the sequence $\mathbf{X}^{(n)}$ to converge to a stationary distribution $\mathbf{X}$. Suppose $\mathbf{X}^{(n)} \to \bar{\mathbf{X}}$ as n goes to infinity then $\bar{\mathbf{X}}$ can be estimated from the sequence $\{X^{(n)}\}$ by computing the proportion of the occurrence of each state in the sequence and let us denote it by $\hat{\mathbf{X}}$. From (6.8) one would expect that

$$\sum_{i=1}^{k} \lambda_i \hat{Q}_i \hat{\mathbf{X}} \approx \hat{\mathbf{X}}. \tag{6.12}$$

This suggests one possible way to estimate the parameters

$$\lambda = (\lambda_1, \ldots, \lambda_k)$$

as follows. One may consider the following minimization problem:

$$\min_{\lambda} || \sum_{i=1}^{k} \lambda_i \hat{Q}_i \hat{\mathbf{X}} - \hat{\mathbf{X}} ||$$

subject to

$$\sum_{i=1}^{k} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.$$

Here $||.||$ is certain vector norm. In particular, if $||.||_\infty$ is chosen, we have the following minimization problem:

$$\min_{\lambda} \max_{l} \left| \left[\sum_{i=1}^{k} \lambda_i \hat{Q}_i \hat{\mathbf{X}} - \hat{\mathbf{X}} \right]_l \right|$$

subject to

$$\sum_{i=1}^{k} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.$$

Here $[\cdot]_l$ denotes the lth entry of the vector. The constraints in the optimization problem guarantee the existence of the stationary distribution $\mathbf{X}$. Next we see that the above minimization problem can be formulated as a linear programming problem:

$$\min_{\lambda} \; w$$

subject to

$$\begin{pmatrix} w \\ w \\ \vdots \\ w \end{pmatrix} \geq \hat{\mathbf{X}} - \left[\hat{Q}_1\hat{\mathbf{X}} \mid \hat{Q}_2\hat{\mathbf{X}} \mid \cdots \mid \hat{Q}_n\hat{\mathbf{X}}\right] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix},$$

$$\begin{pmatrix} w \\ w \\ \vdots \\ w \end{pmatrix} \geq -\hat{\mathbf{X}} + \left[\hat{Q}_1\hat{\mathbf{X}} \mid \hat{Q}_2\hat{\mathbf{X}} \mid \cdots \mid \hat{Q}_n\hat{\mathbf{X}}\right] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix},$$

$$w \geq 0, \quad \sum_{i=1}^{k} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.$$

We can solve the above linear programming problem efficiently and obtain the parameters λ_i. In next subsection, we will demonstrate the estimation method by a simple example.

Instead of solving an min-max problem, one can also choose the $||.||_1$ and formulate the following minimization problem:

$$\min_{\lambda} \sum_{l=1}^{m} \left| \left[\sum_{i=1}^{k} \lambda_i \hat{Q}_i \hat{\mathbf{X}} - \hat{\mathbf{X}} \right]_l \right|$$

subject to

$$\sum_{i=1}^{k} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.$$

The corresponding linear programming problem is given as follows:

$$\min_{\lambda} \sum_{l=1}^{m} w_l$$

subject to

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} \geq \hat{\mathbf{X}} - \left[\hat{Q}_1\hat{\mathbf{X}} \mid \hat{Q}_2\hat{\mathbf{X}} \mid \cdots \mid \hat{Q}_k\hat{\mathbf{X}}\right] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix},$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} \geq -\hat{\mathbf{X}} + \left[\hat{Q}_1\hat{\mathbf{X}} \mid \hat{Q}_2\hat{\mathbf{X}} \mid \cdots \mid \hat{Q}_k\hat{\mathbf{X}}\right] \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix},$$

$$w_i \geq 0, \quad \forall i, \quad \sum_{i=1}^{k} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.$$

In the above linear programming formulation, the number of variables is equal to k and the number of constraints is equal to $(2m+1)$. The complexity of solving a linear programming problem is $O(k^3 L)$ where n is the number of variables and L is the number of binary bits needed to store all the data (the constraints and the objective function) of the problem [91].

We remark that other norms such as $||.||_2$ can also be considered. In this case, it will result in a quadratic programming problem. It is known that in approximating data by a linear function [79, p. 220], $||.||_1$ gives the most robust answer, $||.||_\infty$ avoids gross discrepancies with the data as much as possible and if the errors are known to be normally distributed then $||.||_2$ is the best choice. In the tested examples, we only consider the norms leading to solving linear programming problems.

6.2.3 An Example

We consider a sequence $\{X^{(n)}\}$ of three states $(m = 3)$ given by

$$\{1, 1, 2, 2, 1, 3, 2, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 1, 2\}. \tag{6.13}$$

The sequence $\{X^{(n)}\}$ can be written in vector form

$$X^{(1)} = (1, 0, 0)^T,\ X^{(2)} = (1, 0, 0)^T,\ X^{(3)} = (0, 1, 0)^T,\ \ldots,\ X^{(20)} = (0, 1, 0)^T.$$

We consider $k = 2$, then from (6.13) we have the transition frequency matrices

$$F^{(1)} = \begin{pmatrix} 1 & 3 & 3 \\ 6 & 1 & 1 \\ 1 & 3 & 0 \end{pmatrix} \quad \text{and} \quad F^{(2)} = \begin{pmatrix} 1 & 4 & 1 \\ 3 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix}. \tag{6.14}$$

Therefore from (6.14) we have the i-step transition probability matrices $(i = 1, 2)$ as follows:

$$\hat{Q}_1 = \begin{pmatrix} 1/8 & 3/7 & 3/4 \\ 3/4 & 1/7 & 1/4 \\ 1/8 & 3/7 & 0 \end{pmatrix} \quad \text{and} \quad \hat{Q}_2 = \begin{pmatrix} 1/7 & 4/7 & 1/4 \\ 3/7 & 2/7 & 3/4 \\ 3/7 & 1/7 & 0 \end{pmatrix} \tag{6.15}$$

and

$$\hat{\mathbf{X}} = (\frac{2}{5}, \frac{2}{5}, \frac{1}{5})^T.$$

Hence we have

$$\hat{Q}_1\hat{\mathbf{X}} = (\frac{13}{35}, \frac{57}{140}, \frac{31}{140})^T,$$

and

$$\hat{Q}_2\hat{\mathbf{X}} = (\frac{47}{140}, \frac{61}{140}, \frac{8}{35})^T.$$

To estimate λ_i one can consider the optimization problem:

$$\min_{\lambda_1,\lambda_2} w$$

subject to

$$\begin{cases} w \geq \frac{2}{5} - \frac{13}{35}\lambda_1 - \frac{47}{140}\lambda_2 \\ w \geq -\frac{2}{5} + \frac{13}{35}\lambda_1 + \frac{47}{140}\lambda_2 \\ w \geq \frac{2}{5} - \frac{57}{140}\lambda_1 - \frac{61}{140}\lambda_2 \\ w \geq -\frac{2}{5} + \frac{57}{140}\lambda_1 + \frac{61}{140}\lambda_2 \\ w \geq \frac{1}{5} - \frac{31}{140}\lambda_1 - \frac{8}{35}\lambda_2 \\ w \geq -\frac{1}{5} + \frac{31}{140}\lambda_1 + \frac{8}{35}\lambda_2 \\ w \geq 0, \quad \lambda_1 + \lambda_2 = 1, \quad \lambda_1, \lambda_2 \geq 0. \end{cases}$$

The optimal solution is

$$(\lambda_1^*, \lambda_2^*, w^*) = (1, 0, 0.0286),$$

and we have the model

$$\mathbf{X}^{(n+1)} = \hat{Q}_1\mathbf{X}^{(n)}. \tag{6.16}$$

We remark that if we do not specify the non-negativity of λ_1 and λ_2, the optimal solution becomes

$$(\lambda_1^{**}, \lambda_2^{**}, w^{**}) = (1.80, -0.80, 0.0157),$$

the corresponding model is

$$\mathbf{X}^{(n+1)} = 1.80\hat{Q}_1\mathbf{X}^{(n)} - 0.80\hat{Q}_2\mathbf{X}^{(n-1)}. \tag{6.17}$$

Although w^{**} is less than w^*, the model (6.17) is not suitable. It is easy to check that

$$1.80\hat{Q}_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - 0.80\hat{Q}_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.2321 \\ 1.1214 \\ 0.1107 \end{pmatrix},$$

therefore λ_1^{**} and λ_2^{**} are not valid parameters.

We note that if we consider the minimization problem:

$$\min_{\lambda_1,\lambda_2} w_1 + w_2 + w_3$$

subject to

$$\begin{cases} w_1 \geq \frac{2}{5} - \frac{13}{35}\lambda_1 - \frac{47}{140}\lambda_2 \\ w_1 \geq -\frac{2}{5} + \frac{13}{35}\lambda_1 + \frac{47}{140}\lambda_2 \\ w_2 \geq \frac{2}{5} - \frac{57}{140}\lambda_1 - \frac{61}{140}\lambda_2 \\ w_2 \geq -\frac{2}{5} + \frac{57}{140}\lambda_1 + \frac{61}{140}\lambda_2 \\ w_3 \geq \frac{1}{5} - \frac{31}{140}\lambda_1 - \frac{9}{35}\lambda_2 \\ w_3 \geq -\frac{1}{5} + \frac{31}{140}\lambda_1 + \frac{9}{35}\lambda_2 \\ w_1, w_2, w_3 \geq 0, \quad \lambda_1 + \lambda_2 = 1, \quad \lambda_1, \lambda_2 \geq 0. \end{cases}$$

The optimal solution is the same as the previous min-max formulation and is equal to

$$(\lambda_1^*, \lambda_2^*, w_1^*, w_2^*, w_3^*) = (1, 0, 0.0286, 0.0071, 0.0214).$$

6.3 Some Applications

In this section we apply our model to some data sequences. The data sequences are the DNA sequence and the sales demand data sequence. Given the state vectors $\mathbf{X}^{(i)}$, $i = n-k, n-k+1, \ldots, k-1$, the state probability distribution at time n can be estimated as follows:

$$\hat{\mathbf{X}}^{(n)} = \sum_{i=1}^{k} \lambda_i \hat{Q}_i \mathbf{X}^{(n-i)}.$$

In many applications, one would like to make use of the higher-order Markov chain models for the purpose of prediction. According to this state probability distribution, the prediction of the next state $\hat{X}^{(n)}$ at time n can be taken as the state with the maximum probability, i.e.,

$$\hat{X}^{(n)} = j, \quad \text{if } [\hat{\mathbf{X}}^{(n)}]_i \le [\hat{\mathbf{X}}^{(n)}]_j, \quad \forall 1 \le i \le m.$$

To evaluate the performance and effectiveness of the higher-order Markov chain model, a prediction accuracy r is defined as

$$r = \frac{1}{T} \sum_{t=k+1}^{T} \delta_t,$$

where T is the length of the data sequence and

$$\delta_t = \begin{cases} 1, & \text{if } \hat{X}^{(t)} = X^{(t)} \\ 0, & \text{otherwise.} \end{cases}$$

Using the example in the previous section, two possible prediction rules can be drawn as follows:

$$\begin{cases} \hat{X}^{(n+1)} = 2, & \text{if } X^{(n)} = 1, \\ \hat{X}^{(n+1)} = 1, & \text{if } X^{(n)} = 2, \\ \hat{X}^{(n+1)} = 1, & \text{if } X^{(n)} = 3 \end{cases}$$

or

$$\begin{cases} \hat{X}^{(n+1)} = 2, & \text{if } X^{(n)} = 1, \\ \hat{X}^{(n+1)} = 3, & \text{if } X^{(n)} = 2, \\ \hat{X}^{(n+1)} = 1, & \text{if } X^{(n)} = 3. \end{cases}$$

The prediction accuracy r for the sequence in (6.13) is equal to $12/19$ for both prediction rules. While the prediction accuracies of other rules for the sequence in (6.13) are less than the value $12/19$.

Next we present other numerical results on different data sequences are discussed. In the following tests, we solve min-max optimization problems to determine the parameters λ_i of higher-order Markov chain models. However, we remark that the results of using the $||.||_1$ optimization problem as discussed in the previous section are about the same as that of using the min-max formulation.

6.3.1 The DNA Sequence

In order to determine whether certain short DNA sequence (a categorical data sequence of four possible categories: A,C,G and T) occurred more often than would be expected by chance, Avery [8] examined the Markovian structure of introns from several other genes in mice. Here we apply our model to the introns from the mouse αA-crystallin gene see for instance [175]. We compare our second-order model with the Raftery's second-order model. The model

Table 6.1. Prediction accuracy in the DNA sequence.

	2-state model	3-state model	4-state model
New Model	0.57	0.49	0.33
Raftery's Model	0.57	0.47	0.31
Random Chosen	0.50	0.33	0.25

parameters of the Raftery's model are given in [175]. The results are reported in Table 6.1.

The comparison is made with different grouping of states as suggested in [175]. In grouping states 1 and 3, and states 2 and 4 we have a 2-state model. Our model gives

$$\hat{Q}_1 = \begin{pmatrix} 0.5568 & 0.4182 \\ 0.4432 & 0.5818 \end{pmatrix},$$

$$\hat{Q}_2 = \begin{pmatrix} 0.4550 & 0.5149 \\ 0.5450 & 0.4851 \end{pmatrix}$$

$$\hat{\mathbf{X}} = (0.4858, 0.5142)^T, \quad \lambda_1 = 0.7529 \quad \text{and} \quad \lambda_2 = 0.2471.$$

In grouping states 1 and 3 we have a 3-state model. Our model gives

$$\hat{Q}_1 = \begin{pmatrix} 0.5568 & 0.3573 & 0.4949 \\ 0.2571 & 0.3440 & 0.2795 \\ 0.1861 & 0.2987 & 0.2256 \end{pmatrix},$$

$$\hat{Q}_2 = \begin{pmatrix} 0.4550 & 0.5467 & 0.4747 \\ 0.3286 & 0.2293 & 0.2727 \\ 0.2164 & 0.2240 & 0.2525 \end{pmatrix}$$

$$\hat{\mathbf{X}} = (0.4858, 0.2869, 0.2272)^T, \quad \lambda_1 = 1.0 \quad \text{and} \quad \lambda_2 = 0.0$$

If there is no grouping, we have a 4-state model. Our model gives

$$\hat{Q}_1 = \begin{pmatrix} 0.2268 & 0.2987 & 0.2274 & 0.1919 \\ 0.2492 & 0.3440 & 0.2648 & 0.2795 \\ 0.3450 & 0.0587 & 0.3146 & 0.3030 \\ 0.1789 & 0.2987 & 0.1931 & 0.2256 \end{pmatrix},$$

$$\hat{Q}_2 = \begin{pmatrix} 0.1891 & 0.2907 & 0.2368 & 0.2323 \\ 0.3814 & 0.2293 & 0.2773 & 0.2727 \\ 0.2532 & 0.2560 & 0.2305 & 0.2424 \\ 0.1763 & 0.2240 & 0.2555 & 0.2525 \end{pmatrix}$$