High-dimensional Markov Chain Models for Categorical Data Sequences with Applications

Wai-Ki CHING
AMACL, Department of Mathematics
HKU
The Fifth BIFE 2012, Lanzhou, 19 July 2012

Abstract: Markov chains are popular models for a modelling categorical data sequences. In this talk, I shall present some high-dimensional Markov chain models for modelling categorical data sequences. In particular, both high-order Markov chain models and multivariate Markov chain models will be presented. Estimation methods for the model parameters will be proposed and examples of practical applications in management science will also be given.

Research Supported in Part by HK RGC Grant 7017/07P.
The Outline

(1) Motivations and Objectives.

(2) High-order Markov Chain Models.

(3) Multivariate Markov Chain Models.

(4) Concluding Remarks.
1. Motivations and Objectives.

• Markov chains are popular tools for modeling many practical systems including categorical time series.

• It is also very easy to construct a Markov chain model. Given the observed time series data sequence (Markov chain) \( \{X_t\} \) of \( m \) states, one can count the transition frequency \( F_{jk} \) (one step) in the observed sequence from State \( k \) to State \( j \) in one step.

• Hence one can construct the one-step transition frequency matrix for the observed sequence \( \{X_t\} \) as follows:

\[
F = \begin{pmatrix}
F_{11} & \cdots & \cdots & F_{1m} \\
F_{21} & \cdots & \cdots & F_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
F_{m1} & \cdots & \cdots & F_{mm}
\end{pmatrix}.
\] (1)
• From $F$, one can get the estimates for $P_{kj}$ (column normalization) as follows:

$$
P = \begin{pmatrix}
P_{11} & \cdots & \cdots & P_{1m} \\
P_{21} & \cdots & \cdots & P_{2m} \\
\vdots & \vdots & \vdots & \vdots \\
P_{m1} & \cdots & \cdots & P_{mm}
\end{pmatrix}
$$

(2)

where

$$
P_{kj} = \frac{F_{kj}}{m \sum_{r=1}^{m} F_{rj}}
$$

is the maximum likelihood estimator.
• Example 1: Consider a categorical data sequence of 3 states/categories:
\[ \{1, 1, 2, 2, 1, 3, 2, 1, 2, 3, 1, 2, 3, 1, 2, 1, 2\}. \] (3)

We adopt the following notation. The sequence \( \{X_t\} \) can be written in vector form (canonical representation by Robert J. Elliott):
\[
X_0 = (1, 0, 0)^T, \quad X_1 = (1, 0, 0)^T, \quad X_2 = (0, 1, 0)^T, \ldots, \quad X_{19} = (0, 1, 0)^T.
\]

• The First-order Markov Chain Model:
By counting the transition frequency from State \( k \) to State \( j \) in the sequence, one can construct the transition frequency matrix \( F \) (then the transition probability matrix \( \hat{Q} \)) for the sequence.
\[
F = \begin{pmatrix}
1 & 3 & 3 \\
6 & 1 & 1 \\
1 & 3 & 0
\end{pmatrix}
\text{ and } \hat{Q} = \begin{pmatrix}
1/8 & 3/7 & 3/4 \\
6/8 & 1/7 & 1/4 \\
1/8 & 3/7 & 0
\end{pmatrix} \] (4)

and the first-order Markov chain model is
\[
X_{t+1} = \hat{Q}X_t.
\]
1.1 The Steady-state Probability Distribution

Proposition 1.1: Given an irreducible and aperiodic Markov chain of $m$ states, then for any initial probability distribution $X_0$

$$\lim_{t \to \infty} ||X_t - \pi|| = \lim_{t \to \infty} ||P^t X_0 - \pi|| = 0,$$

where $\pi$ is the steady-state probability distribution of the transition probability matrix $P$ of the underlying Markov chain and $\pi = P \pi$.

- We remark that a non-negative probability vector $\pi$ satisfies $\pi = P \pi$ is called a stationary probability vector. A stationary probability vector is not necessary the steady-state probability distribution vector.
Proposition 1.2: [Perron (1907) - Frobenius (1912) Theorem] Let $A$ be a non-negative, irreducible and aperiodic square matrix of size $m$. Then

(i) $A$ has a positive real eigenvalue $\lambda$, equal to its spectral radius,

$$\lambda = \max_{1 \leq k \leq m} |\lambda_k(A)|$$

where $\lambda_k(A)$ denotes the $k$th eigenvalue of $A$.

(ii) There corresponds an eigenvector $z$, its entries being real and positive, such that $Az = \lambda z$.

(iii) The eigenvalue $\lambda$ is a simple eigenvalue of $A$.

This theorem is important in Markov chains and dynamical systems.
We remark that requirement **aperiodic** is important. The following matrix is non-negative and irreducible but not aperiodic:

\[
Q = \begin{pmatrix}
0 & 0 & \cdots & 0 & 1 \\
1 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & 1 & 0
\end{pmatrix}
\]  
(5)

Now we see that all eigenvalues of \( Q \) are given by

\[
\lambda_j = e^{2\pi i j/n} \quad j = 0, 1, \ldots, n-1
\]

and they all satisfy \( |\lambda_j| = 1 \). They are all on the unit circle. We see that \( \pi = (1, 1, \cdots, 1)^T \) is the **stationary probability vector** but it is not the **steady-state distribution vector**. In fact, begin with \( x_0 = (1, 0, 0, \cdots, 0)^T \), \( x_t \) does not converge to \( \pi \). But we have

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \pi.
\]

- **Categorical data sequences** occur frequently in many real world applications. The delay effects occur in many applications and data.


• “High-order Regime-switching Markov Model for Pricing Exotic Options.” The draft and volatility of the geometric Brownian motion are driven by a high-order Markov chain.

![Figure 1: Simulated Log Return.](image-url)
Figure 2: Prices of Asian (Left) and Lookback (Right) Options Versus Maturity.
2.1 High-order Markov Chain Models.

- High-order Markov chain can better model categorical data sequence (to capture the delay effect).

- Problem: A conventional $n$-th order Markov chain of $m$ states has $O(m^n)$ states and therefore parameters. The number of transition probabilities (to be estimated) increases exponentially with respect to the order $n$ of the model. There is also computational problem.


• Raftery proposed a **high-order Markov chain model** which involves only one additional parameter for **each extra lag**.

• The model:

\[
P(X_t = k \mid X_{t-1} = k_1, \ldots, X_{t-n} = k_n) = \sum_{i=1}^{n} \lambda_i q_{kk_i} \tag{6}
\]

where \( k, k_1, \ldots, k_n \in M \). Here \( M = \{1, 2, \ldots, m\} \) is the set of the possible states and

\[
\sum_{i=1}^{n} \lambda_i = 1 \quad \text{and} \quad Q = [q_{ij}]
\]

is a transition probability matrix with column sums equal to one, such that

\[
0 \leq \sum_{i=1}^{n} \lambda_i q_{kk_i} \leq 1, \quad k, k_1, \ldots, k_n = 1, 2, \ldots, m. \tag{7}
\]
Raftery proved that Model (6) is analogous to the standard AR($n$) model in time series.

The parameters $q_k \lambda_i$ can be estimated numerically by maximizing the log-likelihood of (6) subjected to the constraints (7).

Problems:

(i) This approach involves solving a highly non-linear optimization problem (which is not easy to solve).

(ii) The proposed numerical method neither guarantees convergence nor a global maximum.
2.2 Our High-order Markov Chain Model.

• Raftery’s model can be generalized as follows:

\[ X_{t+n+1} = \sum_{i=1}^{n} \lambda_i Q_i X_{t+n+1-i} \] (8)

• We define \( Q_i \) to be the \( i \)th step transition probability matrix of the sequence.

• We also assume that \( \lambda_i \) are non-negative such that

\[ \sum_{i=1}^{n} \lambda_i = 1 \]

so that the right-hand-side of (8) is a probability distribution.
2.3 A Property of the Model.

Proposition 2.1: If $Q_i$ is irreducible, $\lambda_i > 0$ for $i = 1, 2, \ldots, n$ and
\[
\sum_{i=1}^{n} \lambda_i = 1
\]
then the model in (8) has a stationary distribution $\bar{X}$ when $t \to \infty$
independent of the initial state vectors
\[
X_0, X_1, \ldots, X_{n-1}.
\]
The proof is based on Perron-Frobenius Theorem.

- The stationary distribution $\bar{X}$ is the unique solution of the linear system of equations
\[
(I - \sum_{i=1}^{n} \lambda_i Q_i) \bar{X} = 0 \quad \text{and} \quad 1^T \bar{X} = 1 \tag{9}
\]
where $I$ is the $m$-by-$m$ identity matrix ($m$ is the number of possible states taken by each data point).
2.4 Parameter Estimation.

- Estimation of $Q_i$, the $i$th step transition probability matrix. One can count the transition frequency $f_{jk}^{(i)}$ in the sequence from State $k$ to State $j$ in the $i$th step. We get

$$F^{(i)} = \begin{pmatrix}
  f_{11}^{(i)} & \cdots & f_{m1}^{(i)} \\
  f_{12}^{(i)} & \cdots & f_{m2}^{(i)} \\
  \vdots & \ddots & \vdots \\
  f_{1m}^{(i)} & \cdots & f_{mm}^{(i)}
\end{pmatrix}, \quad \text{for } i = 1, 2, \ldots, n. \quad (10)$$

- From $F^{(i)}$, we get by column normalization:

$$\hat{Q}_i = \begin{pmatrix}
  \hat{q}_{11}^{(i)} & \cdots & \hat{q}_{m1}^{(i)} \\
  \hat{q}_{12}^{(i)} & \cdots & \hat{q}_{m2}^{(i)} \\
  \vdots & \ddots & \vdots \\
  \hat{q}_{1m}^{(i)} & \cdots & \hat{q}_{mm}^{(i)}
\end{pmatrix} \quad \text{where } \hat{q}_{kj}^{(i)} = \frac{f_{kj}^{(i)}}{\sum_{r=1}^{m} f_{rj}^{(i)}} \quad (11)$$
• Linear Programming Formulation for the Estimation of $\lambda_i$.

**Note:** Proposition 2.1 gives a sufficient condition for the sequence $X_t$ to converge to a stationary distribution $\bar{X}$.

• We assume $X_t \rightarrow \bar{X}$ as $t \rightarrow \infty$.

• $\bar{X}$ can be estimated from the sequence $\{X_t\}$ by computing the proportion of the occurrence of each state in the sequence and let us denote it by $\hat{\bar{X}}$.

• From (9) one would expect that

$$\sum_{i=1}^{n} \lambda_i \bar{Q}_i \hat{X} \approx \hat{\bar{X}}. \quad (12)$$
This suggests one possible way to estimate the parameters
\[ \lambda = (\lambda_1, \ldots, \lambda_n). \]

We consider the following optimization problem:
\[
\min_{\lambda} \left\| \sum_{i=1}^{n} \lambda_i \hat{Q}_i \hat{X} - \hat{X} \right\|_\infty = \min_{\lambda} \max_k \left[ \sum_{i=1}^{n} \lambda_i \hat{Q}_i \hat{X} - \hat{X} \right]_k
\]
subject to
\[
\sum_{i=1}^{n} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.
\]

Here \([\cdot]_k\) denotes the \(k\)th entry of the vector.

**Remark:** The optimization problem can be re-formulated as an Linear Programming (LP) problem. Thus it can be solved by using EXCEL for instance.

Remark: Other norms such as $\| \cdot \|_2$ and $\| \cdot \|_1$ can also be considered. The former will result in a quadratic programming problem while $\| \cdot \|_1$ will still result in a linear programming problem.

- It is known that in approximating data by a linear function $\| \cdot \|_1$ gives the most robust answer.

- $\| \cdot \|_\infty$ avoids gross discrepancies with the data as much as possible.

- If the errors are known to be normally distributed then $\| \cdot \|_2$ is the best choice.
The linear programming formulation:

\[
\min_{\lambda} \ w
\]

subject to

\[
\begin{pmatrix}
w \\
\vdots \\
w
\end{pmatrix}
\geq \hat{X} - \begin{bmatrix} Q_1 \hat{X} | Q_2 \hat{X} | \cdots | Q_n \hat{X} \end{bmatrix}\begin{pmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_n
\end{pmatrix},
\]

\[
\begin{pmatrix}
w \\
\vdots \\
w
\end{pmatrix}
\geq -\hat{X} + \begin{bmatrix} Q_1 \hat{X} | Q_2 \hat{X} | \cdots | Q_n \hat{X} \end{bmatrix}\begin{pmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_n
\end{pmatrix},
\]

\[
w \geq 0, \quad \sum_{i=1}^{n} \lambda_i = 1, \quad \text{and} \quad \lambda_i \geq 0, \quad \forall i.
\]

For a numerical demonstration, we refer to Example 1. A 2nd-order \((n = 2)\) model for the 3-states \((m = 3)\) categorical sequence.

We have the transition frequency matrices

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

From (13) we have the \(i\)-step transition probability matrices:

\[
\hat{Q}_1 = \begin{pmatrix}
1/8 & 3/7 & 3/4 \\
3/4 & 1/7 & 1/4 \\
1/8 & 3/7 & 0
\end{pmatrix}
\quad \text{and} \quad
\hat{Q}_2 = \begin{pmatrix}
1/6 & 4/7 & 1/4 \\
1/3 & 2/7 & 3/4 \\
1/2 & 1/7 & 0
\end{pmatrix}
\]  

and the stationary distribution

\[
\hat{X} = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)^T.
\]

Hence we have

\[
\hat{Q}_1\hat{X} = \left(\frac{13}{35}, \frac{57}{140}, \frac{31}{140}\right)^T, \quad \text{and} \quad
\hat{Q}_2\hat{X} = \left(\frac{29}{84}, \frac{167}{420}, \frac{9}{35}\right)^T.
\]
• To estimate $\lambda_i$ we consider the optimization problem:

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

subject to

$$\begin{cases} 
  w \geq \frac{2}{5} - \frac{13}{35} \lambda_1 - \frac{29}{84} \lambda_2 \\
  w \geq -\frac{2}{5} + \frac{13}{35} \lambda_1 + \frac{29}{84} \lambda_2 \\
  w \geq \frac{-2}{5} + \frac{57}{35} \lambda_1 + \frac{84}{167} \lambda_2 \\
  w \geq \frac{2}{5} - \frac{57}{140} \lambda_1 - \frac{420}{167} \lambda_2 \\
  w \geq -\frac{2}{5} + \frac{57}{140} \lambda_1 + \frac{420}{167} \lambda_2 \\
  w \geq \frac{1}{5} + \frac{31}{140} \lambda_1 + \frac{9}{420} \lambda_2 \\
  w \geq \frac{1}{5} - \frac{31}{140} \lambda_1 - \frac{9}{35} \lambda_2 \\
  w \geq -\frac{1}{5} + \frac{31}{140} \lambda_1 + \frac{9}{35} \lambda_2 \\
  w \geq 0, \quad \lambda_1 + \lambda_2 = 1, \quad \lambda_1, \lambda_2 \geq 0.
\end{cases}$$

• The optimal solution is $(\lambda_1^*, \lambda_2^*, w^*) = (1, 0, 0.0286)$. Since $\lambda_2$ is zero, the “optimal” model is a first-order model.
Proposition 2.2 [Zhu and Ching (2011)] If $Q_n$ is irreducible and aperiodic, $\lambda_1, \lambda_n > 0$ and 

$$\sum_{i=1}^{n} \lambda_i = 1$$

then the model has a stationary distribution $\mathbf{X}$ satisfying

$$(I - \sum_{i=1}^{n} \lambda_i Q_i)\bar{X} = 0 \quad \text{and} \quad 1^T\bar{X} = 1$$

and

$$\lim_{t \to \infty} \mathbf{X}_t = \mathbf{X}.$$ 

- We remark that if $\lambda_n = 0$ then it is not an $n$th model and if $\lambda_1 = 0$ then the model is clearly reducible.
2.5 The Newsboy’s Problem.

- A newsboy sells newspaper (perishable product) every morning. The cost of each newspaper remains at the end of the day is $C_o$ (overage cost) and the cost of each unsatisfied demand is $C_s$ (shortage cost).

- Suppose that the (stationary distribution) probability density function of the demand $D$ is given by

$$\text{Prob} (D = d) = p_d \geq 0, \quad d = 1, 2, \ldots, m.$$  \hspace{1cm} (15)

- To determine the best amount $r^*$ (order size) of newspaper to be ordered such that the expected cost is minimized.

**Proposition 2.2:** The optimal order size $r^*$ is the one which satisfies

$$F(r^* - 1) < \frac{C_s}{C_s + C_o} \leq F(r^*).$$  \hspace{1cm} (16)

Here $F(x) = \sum_{i=1}^{x} p_i$. 


Example 2: Suppose that the demand \((1, 2, \ldots, 2k)\) \((m = 2k)\) follows an Markov process with the transition probability matrix \(Q\) of size \(2k \times 2k\) given by

\[
Q = \begin{pmatrix}
0 & 0 & \cdots & 0 & 1 \\
1 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & : \\
: & \ddots & \ddots & \ddots & : \\
0 & \cdots & 0 & 1 & 0
\end{pmatrix}
\] (17)

- Assume that \(C_o = C_s\). Clearly the next “demand” can be determined by the state of the current demand, and hence the optimal expected cost is equal to zero when the first-order Markov model is used.

- When the classical Newsboy model is used, we note that the stationary distribution of \(Q\) is given by

\[
\frac{1}{2k} (1, 1, \ldots, 1)^T.
\]

The optimal ordering size is equal to \(k\) by Proposition 2.2 and therefore the optimal expected cost is \(C_o k\).
**Example 3:** A large soft-drink company in Hong Kong faces an in-house problem of production planning and inventory control.

- Products are labeled as either **very high sales volume** (state 1), **high sales volume** (state 2), **standard sales volume** (state 3), **low sales volume** (state 4) or **very low sales volume** (state 5). ($m = 5$).

- For simplicity, we assume the following **symmetric cost matrix**:

\[
C = \begin{pmatrix}
0 & 100 & 300 & 700 & 1500 \\
100 & 0 & 100 & 300 & 700 \\
300 & 100 & 0 & 100 & 300 \\
700 & 300 & 100 & 0 & 100 \\
1500 & 700 & 300 & 100 & 0
\end{pmatrix}
\]  

where $[C]_{ij}$ is the assigned **penalty cost** when the production plan is for sales volume of State $i$ and the actual sales volume is State $j$. 

27
• We employ high-order Markov model for modeling the sales demand data sequence.

• Optimal production policy can then be derived based on Proposition 2.2.

• The following table shows the optimal costs based on three different models for three different products.

<table>
<thead>
<tr>
<th>Products</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-order Markov Model ($n = 3$)</td>
<td>11200</td>
<td>9300</td>
<td>10800</td>
</tr>
<tr>
<td>First-order Markov Model ($n = 1$)</td>
<td>27600</td>
<td>18900</td>
<td>11100</td>
</tr>
<tr>
<td>Stationary Model</td>
<td>31900</td>
<td>18900</td>
<td>16300</td>
</tr>
</tbody>
</table>


- In many situations, there is a need to consider a number of categorical data sequences (having same number of categories/states) together at the same time because they are related to each other.


- **Problem:** The conventional first-order Markov chain model for $s$ categorical data sequences of $m$ states has $m^s$ states.
3.1 The Multivariate Markov Chain Model.

- We propose a multivariate Markov chain model which can capture both the **intra- and inter-transition** probabilities among the sequences and the number of model parameters is $O(s^2m^2)$.

- Given $s$ sequences with fixed $m$, the minimum number of parameters required is $\binom{s}{2} = O(s^2)$.

- We assume that there are $s$ categorical sequences and each has $m$ possible states in $M = \{1, 2, \ldots, m\}$.

- Let $X_{n}^{(k)}$ be the state probability distribution vector of the $k$th sequence at time $n$. If the $k$th sequence is in State $j$ at time $n$ then
  \[ X_{n}^{(k)} = e_j = (0, \ldots, 0, \overset{1}{\underset{j\text{th entry}}{\underbrace{1}}}, 0 \ldots, 0)^T. \]
• In our proposed multivariate Markov chain model, we assume the following relationship:

\[
X_{n+1}^{(j)} = \sum_{k=1}^{s} \lambda_{jk} P^{(jk)} X_n^{(k)}, \quad \text{for} \quad j = 1, 2, \cdots, s
\]  

(19)

where

\[
\lambda_{jk} \geq 0, \quad 1 \leq j, k \leq s \quad \text{and} \quad \sum_{k=1}^{s} \lambda_{jk} = 1, \quad \text{for} \quad j = 1, 2, \cdots, s.
\]  

(20)

• The state probability distribution of the \textit{kth sequence} at time \((n + 1)\) depends on the weighted average of \(P^{(jk)} X_n^{(k)}\).

• Here \(P^{(jk)}\) is a one-step transition probability matrix from the states in the \textit{kth sequence} to the states in the \textit{jth sequence}. Here recall that \(X_n^{(k)}\) is the state probability distribution of the \textit{kth sequences} at time \(n\).
• In matrix form, we have the following block structure matrix equation (a compact representation):

\[ X_{n+1} = \begin{pmatrix} X^{(1)}_{n+1} \\ X^{(2)}_{n+1} \\ \vdots \\ X^{(s)}_{n+1} \end{pmatrix} = \begin{pmatrix} \lambda_{11} P^{(11)} & \lambda_{12} P^{(12)} & \cdots & \lambda_{1s} P^{(1s)} \\ \lambda_{21} P^{(21)} & \lambda_{22} P^{(22)} & \cdots & \lambda_{2s} P^{(2s)} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{s1} P^{(s1)} & \lambda_{s2} P^{(s2)} & \cdots & \lambda_{ss} P^{(ss)} \end{pmatrix} \begin{pmatrix} X^{(1)}_n \\ X^{(2)}_n \\ \vdots \\ X^{(s)}_n \end{pmatrix} \]

or

\[ X_{n+1} = QX_n. \]

Remark: We note that \( X^{(j)}_n \) is a probability distribution vector.

Remark: A research problem is the following. If \( Q \) is given, is there any efficient way to get \( X \) such that \( X = QX \)?
### 3.2 Some Properties of the Model.

**Proposition 3.1**: If $\lambda_{jk} > 0$ for $1 \leq j, k \leq s$, then the matrix $Q$ has an eigenvalue equal to one and the eigenvalues of $Q$ have modulus less than or equal to one. [Compare with the Proposition 1.2.]

**Proposition 3.2**: Suppose that $P^{(jk)} (1 \leq j, k \leq s)$ are irreducible and $\lambda_{jk} > 0$ for $1 \leq j, k \leq s$. Then there is a vector $X = [X^{(1)}, \ldots, X^{(s)}]^T$ such that

\[
X = QX \quad \text{and} \quad \sum_{i=1}^{m} [X^{(j)}]_i = 1, \quad 1 \leq j \leq s.
\]

Moreover,

\[
\lim_{n \to \infty} X_n = X.
\]

Again the proof depends on Perron-Frobenius theorem.

**Remark**: Propositions 3.1 and 3.2 still hold if $[\Lambda_{ij}]$ is irreducible and aperiodic [Compare with Propositions 1.1 and 1.2].
3.3 Parameter Estimation.

3.3.1 Estimations of $P^{(jk)}$.

- We can construct the transition frequency matrix from the observed data sequences. More precisely, we count the transition frequency $f^{(jk)}_{i_ji_k}$ from the state $i_k$ in the sequence $\{X_n^{(k)}\}$ to the state $i_j$ in the sequence $\{X_n^{(j)}\}$.

- Therefore we construct the transition frequency matrix for the sequences as follows:

$$F^{(jk)} = \begin{pmatrix}
  f^{(jk)}_{11} & \cdots & \cdots & f^{(jk)}_{m1} \\
  f^{(jk)}_{12} & \cdots & \cdots & f^{(jk)}_{m2} \\
  \vdots & \ddots & \vdots & \vdots \\
  f^{(jk)}_{1m} & \cdots & \cdots & f^{(jk)}_{mm}
\end{pmatrix}.$$
From $F^{(jk)}$, we get the estimates for $P^{(jk)}$ as follows:

$$\hat{P}^{(jk)} = \begin{pmatrix}
\hat{p}_{11}^{(jk)} & \cdots & \cdots & \hat{p}_{m1}^{(jk)} \\
\hat{p}_{12}^{(jk)} & \cdots & \cdots & \hat{p}_{m2}^{(jk)} \\
\vdots & \ddots & \ddots & \vdots \\
\hat{p}_{1m}^{(jk)} & \cdots & \cdots & \hat{p}_{mm}^{(jk)}
\end{pmatrix}$$

where

$$\hat{p}_{i_ji_k}^{(jk)} = \frac{f_{i_ji_k}^{(jk)}}{\sum_{i_j=1}^{m} f_{i_ji_k}^{(jk)}}$$
3.3.2 Estimation of $\lambda_{jk}$

- We have seen that the multivariate Markov chain has a stationary vector (joint probability distribution vector) $X$.

- The vector $X$ can be estimated from the sequences by computing the proportion of the occurrence of each state in each of the sequences, and let us denote it by

$$\hat{X} = (\hat{X}^{(1)}, \hat{X}^{(2)}, \ldots, \hat{X}^{(s)})^T.$$  

One would expect that

$$\begin{pmatrix}
\hat{X}^{(1)} \\
\hat{X}^{(2)} \\
\vdots \\
\hat{X}^{(s)}
\end{pmatrix}
\approx
\begin{pmatrix}
\hat{P}^{(11)} & \hat{P}^{(12)} & \ldots & \hat{P}^{(1s)} \\
\hat{P}^{(21)} & \hat{P}^{(22)} & \ldots & \hat{P}^{(2s)} \\
\vdots & \vdots & \ddots & \vdots \\
\hat{P}^{(s1)} & \hat{P}^{(s2)} & \ldots & \hat{P}^{(ss)}
\end{pmatrix} \hat{X}.$$  

(21)
• If $\|\cdot\|_\infty$ is chosen to minimize the discrepancies then we have the following optimization problem:

$$\begin{align*}
\min_{\lambda} \left\{ \max_i \left[ \sum_{k=1}^{m} \lambda_{jk} \hat{P}(jk) \hat{X}(k) - \hat{X}(j) \right] \right\} \\
\text{subject to} \sum_{k=1}^{s} \lambda_{jk} = 1, \quad \text{and} \quad \lambda_{jk} \geq 0, \quad \forall k.
\end{align*}$$

(22)

**Remark:** Again other norms such as $\|\cdot\|_2$ and $\|\cdot\|_1$ can also be considered. The former will result in a **quadratic programming problem** while $\|\cdot\|_1$ will still result in a **linear programming problem**.
Problem (22) can be formulated as \textit{s linear programming problems}. For each \( j \):

\[
\min_{\lambda} w_j
\]

subject to

\[
\begin{align*}
\begin{bmatrix}
w_j \\
w_j \\
\vdots \\
w_j \\
w_j
\end{bmatrix} & \geq \hat{x}(j) - B_j \\
\begin{bmatrix}
w_j \\
w_j \\
\vdots \\
w_j
\end{bmatrix} & \geq -\hat{x}(j) + B_j,
\end{align*}
\]

\[
\begin{align*}
\sum_{k=1}^{s} \lambda_{jk} &= 1, \\
\lambda_{jk} &\geq 0, \quad \forall k,
\end{align*}
\]

where

\[
B_j = [\hat{p}(j1)\hat{X}(1) \mid \hat{p}(j2)\hat{X}(2) \mid \cdots \mid \hat{p}(js)\hat{X}(s)].
\]
3.4 A Numerical Demonstration

• Consider \( s = 2 \) sequences of \( m = 4 \) states:

\[
S_1 = \{4, 3, 1, 3, 4, 4, 3, 3, 1, 2, 3, 4\}
\]

and

\[
S_2 = \{1, 2, 3, 4, 1, 4, 4, 3, 3, 1, 3, 1\}.
\]

• By counting the transition frequencies

\[
S_1 : 4 \rightarrow 3 \rightarrow 1 \rightarrow 3 \rightarrow 4 \rightarrow 4 \rightarrow 3 \rightarrow 3 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4
\]

and

\[
S_2 : 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1 \rightarrow 4 \rightarrow 4 \rightarrow 3 \rightarrow 3 \rightarrow 1 \rightarrow 3 \rightarrow 1
\]

we have

\[
F^{(11)} = \begin{pmatrix}
0 & 0 & 2 & 0 \\
1 & 0 & 0 & 0 \\
1 & 1 & 1 & 2 \\
0 & 0 & 2 & 1
\end{pmatrix}
\]

and

\[
F^{(22)} = \begin{pmatrix}
0 & 0 & 2 & 1 \\
1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1
\end{pmatrix}.
\]
Moreover by counting the inter-transition frequencies

\[
S_1 : 4 \quad 3 \quad 1 \quad 3 \quad 4 \quad 4 \quad 3 \quad 3 \quad 1 \quad 2 \quad 3 \quad 4
\]

\[
S_2 : 1 \quad 2 \quad 3 \quad 4 \quad 1 \quad 4 \quad 4 \quad 3 \quad 3 \quad 1 \quad 3 \quad 1
\]

and

\[
S_1 : 4 \quad 3 \quad 1 \quad 3 \quad 4 \quad 4 \quad 3 \quad 3 \quad 1 \quad 2 \quad 3 \quad 4
\]

\[
S_2 : 1 \quad 2 \quad 3 \quad 4 \quad 1 \quad 4 \quad 4 \quad 3 \quad 3 \quad 1 \quad 3 \quad 1
\]

We have

\[
F^{(21)} = \begin{pmatrix}
1 & 0 & 2 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 3 & 0 \\
1 & 0 & 0 & 2
\end{pmatrix}, \quad F^{(12)} = \begin{pmatrix}
0 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 \\
2 & 0 & 1 & 2 \\
1 & 0 & 1 & 1
\end{pmatrix}.
\]
After normalization we have the transition probability matrices:

\[ \hat{P}^{(11)} = \begin{pmatrix} 0 & 0 & \frac{2}{5} & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 1 & \frac{1}{5} & \frac{2}{3} \\ 0 & 0 & \frac{2}{5} & \frac{1}{3} \end{pmatrix}, \quad \hat{P}^{(12)} = \begin{pmatrix} 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ \frac{2}{3} & 0 & \frac{1}{4} & \frac{2}{3} \\ \frac{1}{3} & 0 & \frac{1}{4} & \frac{1}{3} \end{pmatrix}, \]

\[ \hat{P}^{(21)} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{2}{5} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \\ 0 & 1 & \frac{3}{5} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{2}{3} \end{pmatrix}, \quad \hat{P}^{(22)} = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{12} & 1 & \frac{1}{4} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{4} & \frac{1}{3} \end{pmatrix}. \]

Moreover, we also have

\[ \hat{X}_1 = \left( \frac{1}{6}, \frac{1}{12}, \frac{5}{12}, \frac{1}{3} \right)^T \quad \text{and} \quad \hat{X}_2 = \left( \frac{1}{3}, \frac{1}{12}, \frac{1}{3}, \frac{1}{4} \right)^T \]

Solving the corresponding linear programming problems, the multivariate Markov models of the two categorical data sequences \( S_1 \) and \( S_2 \) are then given by

\[ \begin{cases} 
X_n^{(1)} & = 0.00 \hat{P}^{(11)} X_n^{(1)} + 1.00 \hat{P}^{(12)} X_n^{(2)} \\
X_n^{(2)} & = 0.89 \hat{P}^{(21)} X_n^{(1)} + 0.11 \hat{P}^{(22)} X_n^{(2)}.
\end{cases} \]
Proposition 3.3 [Zhu and Ching (2011)] If $\lambda_{ii} > 0$, $P_{ii}$ is irreducible (for $1 \leq i \leq s$), the matrix $[\lambda_{ij}]$ is also irreducible and at least one of $P_{ii}$ is aperiodic then the model has a stationary joint probability distribution

$$X = (x^{(1)}, \ldots, x^{(s)})^T$$

satisfying

$$X = QX.$$ 

Moreover, we have

$$\lim_{t \to \infty} X_t = X.$$ 

3.5 A Simplified Model.

\[
x_{n+1} \equiv \begin{pmatrix}
    x_{n+1}^{(1)} \\
x_{n+1}^{(2)} \\
    \vdots \\
x_{n+1}^{(s)}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
    \lambda_{11} P^{(11)} & \lambda_{12} I & \cdots & \lambda_{1s} I \\
    \lambda_{21} I & \lambda_{22} P^{(22)} & \cdots & \lambda_{2s} I \\
    \vdots & \vdots & \ddots & \vdots \\
    \lambda_{s1} I & \lambda_{s2} I & \cdots & \lambda_{ss} P^{(ss)}
\end{pmatrix}
\begin{pmatrix}
    x_n^{(1)} \\
x_n^{(2)} \\
    \vdots \\
x_n^{(s)}
\end{pmatrix}
\]

\[
\equiv Qx_n
\]  

- The models can still capture both the intra- and inter-transition probabilities among the sequences but the number of parameters in the proposed model is \(O(sm^2 + s^2)\).
Proposition 3.4. Suppose that $P^{(jj)} \ (1 \leq j \leq s)$ and $\Lambda = [\lambda_{jk}]^T$ are irreducible. Then there is a vector

$$x = (x^{(1)}, x^{(2)}, \ldots, x^{(s)})^T$$

such that $x = Qx$ and

$$\sum_{i=1}^{m} [x^{(j)}]_i = 1, \ 1 \leq j \leq s.$$ 

• The estimation methods described previously for the model parameters $\lambda_{ij}$ and $P_{ij}$ can still be applied.

3.6 Further Numerical Examples

● A soft-drink company in Hong Kong presently faces an in-house problem of production planning and inventory control. A pressing issue that stands out is the storage space of its central warehouse, which often finds itself in the state of overflow or near capacity.

● Products (A,B,C,D,E) at different time \( t \) can be categorized into five possible states according to sales volume. All products are labelled as either
  (1) very fast-moving (very high sales volume),
  (2) fast-moving,
  (3) standard,
  (4) slow-moving, or
  (5) very slow-moving (very low sales volume).
• We then build the multivariate Markov model, to predict the next state $\hat{x}_t$ at time $t$ which can be taken as the state with the maximum probability, i.e.,

$$\hat{x}_t = j, \text{ if } [\hat{x}_t]_i \leq [\hat{x}_t]_j, \forall 1 \leq i \leq m.$$ 

• To evaluate the performance and effectiveness of our multivariate Markov chain model, a prediction result is measured by the prediction accuracy $r$ defined as

$$r = \frac{1}{T} \times \sum_{t=1}^{T} \delta_t \times 100\%,$$

where $T$ is the length of the data sequence and

$$\delta_t = \begin{cases} 
1, & \text{if } \hat{x}_t = x_t \\
0, & \text{otherwise.}
\end{cases}$$
Another way to compare the performance of the models is to use the BIC (Bayesian Information Criterion) which is defined as

\[ BIC = -2L + q \log n, \]

where

\[ L = \sum_{j=1}^{s} \sum_{i_0, k_1, \ldots, k_s=1}^{m} n_{i_0, k_1, \ldots, k_s}^{(j)} \log \left( \sum_{l=1}^{m} \sum_{k=1}^{s} \lambda_{jk} p_{i_0, k_l}^{(jk)} \right), \]

is the log-likelihood of the model,

\[ n_{i_0, k_1, k_2, \ldots, k_s}^{(j)} = \sum x_{n+1}^{(j)} (i_0) x_{n}^{1}(k_1) x_{n}^{2}(k_2) \cdots x_{n}^{s}(k_s), \]

and \( q \) is the number of independent parameters, and \( n \) is the length of the sequence.

- The smaller the value of BIC, the better the model is.
Models

Multivariate Markov Model
Simplified Model

<table>
<thead>
<tr>
<th>Models</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Markov Model</td>
<td>50%</td>
<td>45%</td>
<td>63%</td>
<td>52%</td>
<td>55%</td>
</tr>
<tr>
<td>Simplified Model</td>
<td>46%</td>
<td>46%</td>
<td>63%</td>
<td>52%</td>
<td>54%</td>
</tr>
</tbody>
</table>

Table 1. Prediction Accuracy for the Sales Demand Data Sequences of Full Length.

<table>
<thead>
<tr>
<th>Models</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Markov Model</td>
<td>8.0215e+003</td>
</tr>
<tr>
<td>Simplified Model</td>
<td>3.9878e+003</td>
</tr>
</tbody>
</table>

Table 2. BIC for the Sales Demand Data Sequences in Full Length.
### Table 3. Prediction Accuracy for the Sales Demand Data Sequences of Shorter Length of 30.

<table>
<thead>
<tr>
<th>Models</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Markov Model</td>
<td>66%</td>
<td>59%</td>
<td>83%</td>
<td>55%</td>
<td>55%</td>
</tr>
<tr>
<td>Simplified Model</td>
<td>66%</td>
<td>55%</td>
<td>83%</td>
<td>55%</td>
<td>52%</td>
</tr>
</tbody>
</table>

### Table 4. BIC for the Sales Demand Data Sequences in Shorter Length of 30.

<table>
<thead>
<tr>
<th>Models</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariate Markov Model</td>
<td>3.3714e+003</td>
</tr>
<tr>
<td>Simplified Model</td>
<td>934.3750</td>
</tr>
</tbody>
</table>

4.1 Extension to High-order Multivariate Markov Chain Models

- We assume that there are \( s \) categorical sequences with order \( n \) and each has \( m \) possible states in \( M \). In the proposed model, we assume that the \( j \)th sequence at time \( t = r + 1 \) depends on all the sequences at times \( t = r, r - 1, \ldots, r - n + 1 \).

- Using the same notations, our proposed high-order \((n)th-order\) multivariate Markov chain model takes the following form:

\[
\mathbf{x}_{r+1}^{(j)} = \sum_{k=1}^{s} \sum_{h=1}^{n} \lambda_{jk}^{(h)} \mathbf{P}_h^{(j)} \mathbf{x}_{r-h+1}^{(k)}, \quad j = 1, 2, \ldots, s \tag{27}
\]

where

\[
\lambda_{jk}^{(h)} \geq 0, \quad 1 \leq j, k \leq s, \quad 1 \leq h \leq n \quad \text{and} \quad \sum_{k=1}^{s} \sum_{h=1}^{n} \lambda_{jk}^{(h)} = 1, \quad j = 1, 2, \ldots, s.
\]

In fact, if we let

\[ \mathbf{X}_r^{(j)} = ((\mathbf{x}_r^{(j)})^T, (\mathbf{x}_{r-1}^{(j)})^T, \ldots, (\mathbf{x}_{r-n+1}^{(j)})^T)^T \]

for \( j = 1, 2, \ldots, s \) be the \( nm \times 1 \) vectors then the model can be written as the following matrix form:

\[
\begin{pmatrix}
\mathbf{X}_r^{(1)} \\
\mathbf{X}_r^{(2)} \\
\vdots \\
\mathbf{X}_r^{(s)}
\end{pmatrix}
= 
\begin{pmatrix}
\mathbf{B}^{(11)} & \mathbf{B}^{(12)} & \cdots & \mathbf{B}^{(1s)} \\
\mathbf{B}^{(21)} & \mathbf{B}^{(22)} & \cdots & \mathbf{B}^{(2s)} \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{B}^{(s1)} & \mathbf{B}^{(s2)} & \cdots & \mathbf{B}^{(ss)}
\end{pmatrix}
\begin{pmatrix}
\mathbf{X}_r^{(1)} \\
\mathbf{X}_r^{(2)} \\
\vdots \\
\mathbf{X}_r^{(s)}
\end{pmatrix}
\equiv \mathbf{JX}_r
\]
where

\[
B^{(ii)} = \begin{pmatrix}
\lambda_{ii}^{(1)} P_1^{(ii)} & \lambda_{ii}^{(2)} P_2^{(ii)} & \ldots & \lambda_{ii}^{(n-1)} P_{n-1}^{(ii)} & \lambda_{ii}^{(n)} P_n^{(ii)} \\
I & 0 & \ldots & 0 & 0 \\
0 & I & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \ldots & 0 & I & 0
\end{pmatrix}_{mn \times mn}
\]

and if \( i \neq j \) then

\[
B^{(ij)} = \begin{pmatrix}
\lambda_{ij}^{(1)} P_1^{(ij)} & \lambda_{ij}^{(2)} P_2^{(ij)} & \ldots & \lambda_{ij}^{(n-1)} P_{n-1}^{(ij)} & \lambda_{ij}^{(n)} P_n^{(ij)} \\
0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & \ldots & 0 & 0 & 0
\end{pmatrix}_{mn \times mn}
\]

**Remark:** A research problem, if given \( J \), any efficient method to get the stationary vector?
• We define an \( s \times s \) matrix \( \tilde{B} \) first.

We let \( \tilde{B}_{ii} = 1 \) for all \( i = 1, 2, \ldots, s \), and if \( i \neq j \)

\[
\tilde{B}_{ij} = \begin{cases} 
1 & \text{if } \lambda_{ij}^{(k)} > 0 \text{ for all } k \\
0 & \text{otherwise.}
\end{cases}
\]

Here \( j = 1, 2, \ldots, s, k = 1, 2, \ldots, n \).

Proposition 4.1 If \( \lambda_{ii}^{(1)}, \lambda_{ii}^{(n)} > 0 \), \( P_1^{(ii)} \) is irreducible and at least one of them is aperiodic \( (i = 1, 2, \ldots, s) \), additionally, \( \tilde{B} \) is irreducible, then the model has a stationary probability distribution \( \mathbf{X} \) satisfying

\[ \mathbf{X} = J\mathbf{X}. \]

Moreover,

\[ \lim_{t \to \infty} \mathbf{X}_t = \mathbf{X}. \]
4.2 Extension to Negative $\lambda_{ij}$

• Theoretical results of the multivariate Markov chains were obtained when $\lambda_{ij} \geq 0$ (non-negative matrix). It is interesting to study the properties of the models when $\lambda_{ij}$ are allowed to be negative.

• An example of a two-chain model:

$$
\begin{pmatrix}
  x^{(1)}_{n+1} \\
  x^{(2)}_{n+1}
\end{pmatrix} =
\begin{pmatrix}
  \lambda_{1,1}P^{(11)} & \lambda_{1,2}I \\
  \lambda_{2,1}I & \lambda_{2,2}P^{(22)}
\end{pmatrix}
\begin{pmatrix}
  x^{(1)}_n \\
  x^{(2)}_n
\end{pmatrix}
\underbrace{+ \frac{1}{m-1} \begin{pmatrix}
  \lambda_{1,-1}P^{(11)} & \lambda_{1,-2}I \\
  \lambda_{2,-1}I & \lambda_{2,-2}P^{(22)}
\end{pmatrix} \begin{pmatrix}
  1 - x^{(1)}_n \\
  1 - x^{(2)}_n
\end{pmatrix}}_{\text{Negative correlated part}}.
$$

Here $\lambda_{i,j} \geq 0$ for $i = 1, 2$ and $j = \pm 1, \pm 2$ and $\sum_{j=-2}^{2} \lambda_{i,j} = 1$.

Wai-Ki Ching and Michael K. Ng
Markov Chains : Models, Algorithms and Applications