Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 3, No. 2, pp. 235 – 243, May 2007

ON MULTI-DIMENSIONAL MARKOV CHAIN MODELS*

Wai-Ki Ching, Shu-Qin Zhang and Michael K. Ng

Abstract: Markov chain models are commonly used to model categorical data sequences. In this paper, we propose a multi-dimensional Markov chain model for modeling high dimensional categorical data sequences. In particular, the model is practical when there are limited data available. We then test the model with some practical sales demand data. Numerical results indicate the proposed model when compared to the existing models has comparable performance but has much less number of model parameters.

Key words: *high dimensional Markov chains, categorical data sequences, demand prediction*

Mathematics Subject Classification: *65C20, 65F10*

1 Introduction

Categorical data sequences have many applications in both applied sciences and engineering sciences such as inventory control [3, 4, 5], data mining [7], bioinformatics [6] and many other applications [11]. Very often, one has to consider multiple Markov chains (data sequences) at the same time. This is because very often the chains (sequences) can be correlated and therefore the information of other chains can contribute to the chain considered. Thus by exploring these relationships, one can develop better models for better prediction rules. We note that the conventional Markov chain model for s categorical data sequences of m states has m^s states. It is a high dimensional Markov chain process. The number of parameters (transition probabilities) increases exponentially with respect to the number of categorical sequences. This huge number of parameters discourages people from using such kind of Markov chain models. In view of this, Ching et al. proposed a first-order multivariate Markov chain model in [4] for this concerned problem. They then applied the model to the prediction of sales demands of multiple products. Their model involves $O(s^2m^2 + s^2)$ number of parameters where s is the number of sequences and m is the number of possible states. They also developed efficient estimation methods for the model parameters. In [6], the multivariate Markov chain model was then used in building stochastic networks for gene expression sequences. An application of the multivariate Markov chain model to modelling credit risk has been also discussed in [14].

In this paper, we propose simplified multivariate Markov models based on [4] for modelling multiple categorical data sequences. The models can capture both the intra- and inter-transition probabilities among the sequences but the number of parameters is only

*Research supported in part by Hung Hing Ying Physical Research Fund, HKU GRCC Grants Nos. 10206647, 10206483 and 10206147.

$O(sm^2 + s^2)$. We also develop parameter estimation methods based on linear programming for estimating the model parameters. We then apply the model and the method to sales demand data sequences. Numerical results indicate that the new models have good prediction accuracy when compared to the model in [4].

The rest of the paper is organized as follows. In Section 2, we propose a new simplified multivariate Markov model and discuss some important properties of the model. In Section 3, we present the method for the estimation of model parameters. In Section 4, we apply the new simplified model and the numerical method to the sales demand data. In Section 5, we discuss further modification of the model for the case when the observed sequences are very short. Finally, concluding remarks are given in Section 6 to address further research issues.

2 The Multivariate Markov Chain Model

In this section, we first propose our new simplified multivariate Markov chain model and then some of its properties. In the new multivariate Markov chain model, we assume that there are s categorical sequences and each has m possible states in M. We also adopt the following notations. Let $\mathbf{x}_n^{(k)}$ be the state probability distribution vector of the kth Sequence at time n. If the kth Sequence is in State j with probability one at time n then we write

$$\mathbf{x}_n^{(k)} = \mathbf{e}_j = (0, \ldots, 0, \underbrace{1}_{j\text{th entry}}, 0 \ldots, 0)^T.$$

Moreover, we assume the following relationship among the sequences:

$$\mathbf{x}_{n+1}^{(j)} = \lambda_{jj} P^{(jj)} \mathbf{x}_n^{(j)} + \sum_{k=1, k\neq j}^{s} \lambda_{jk} \mathbf{x}_n^{(k)}, \quad \text{for} \quad j = 1, 2, \ldots, s \tag{2.1}$$

where

$$\lambda_{jk} \geq 0, \quad 1 \leq j, k \leq s \quad \text{and} \quad \sum_{k=1}^{s} \lambda_{jk} = 1, \quad \text{for} \quad j = 1, 2, \ldots, s. \tag{2.2}$$

Equation (2.1) simply means that the state probability distribution of the jth chain at time $(n+1)$ depends only on the weighted average of $P^{(jj)}\mathbf{x}_n^{(j)}$ and the state probability distribution of other chains at time n. Here $P^{(jj)}$ is the one-step transition probability matrix of the jth Sequence. In matrix form, one may write

$$\begin{aligned} \mathbf{x}_{n+1} \equiv \begin{pmatrix} \mathbf{x}_{n+1}^{(1)} \\ \mathbf{x}_{n+1}^{(2)} \\ \vdots \\ \mathbf{x}_{n+1}^{(s)} \end{pmatrix} &= \begin{pmatrix} \lambda_{11}P^{(11)} & \lambda_{12}I & \cdots & \lambda_{1s}I \\ \lambda_{21}I & \lambda_{22}P^{(22)} & \cdots & \lambda_{2s}I \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{s1}I & \lambda_{s2}I & \cdots & \lambda_{ss}P^{(ss)} \end{pmatrix} \begin{pmatrix} \mathbf{x}_n^{(1)} \\ \mathbf{x}_n^{(2)} \\ \vdots \\ \mathbf{x}_n^{(s)} \end{pmatrix} \\ &\equiv Q\mathbf{x}_n \end{aligned} \tag{2.3}$$

For Model (2.3), we have the following proposition which can be considered as a generalized version of the Perron-Frobenius Theorem [10, pp. 508-511].

Theorem 2.1 (Perron-Frobenius Theorem). *Let A be a nonnegative and irreducible square matrix of order m. Then*

(i) *A has a positive real eigenvalue, λ, equal to its spectral radius, i.e. $\lambda = \max_k |\lambda_k(A)|$ where $\lambda_k(A)$ denotes the kth eigenvalue of A.*

(ii) *There corresponds an eigenvector $\mathbf{z}$, its entries being real and positive, such that $A\mathbf{z} = \lambda\mathbf{z}$.*

(iii) *λ is a simple eigenvalue of A.*

Proposition 2.2. *Suppose that $P^{(jj)}$ $(1 \le j \le s)$ and $\Lambda = [\lambda_{jk}]^T$ are irreducible. Then there is a vector*

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

such that $\mathbf{x} = Q\mathbf{x}$ and

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

Proof. From (2.2), each column sum of the following matrix

$$\Lambda = \begin{pmatrix} \lambda_{1,1} & \lambda_{2,1} & \cdots & \lambda_{s,1} \\ \lambda_{1,2} & \lambda_{2,2} & \cdots & \lambda_{s,2} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{1,s} & \lambda_{2,s} & \cdots & \lambda_{s,s} \end{pmatrix}$$

is equal to one. Since Λ is nonnegative and irreducible, from the Perron-Frobenius Theorem, there exists a vector

$$\mathbf{y} = (y_1, y_2, \ldots, y_s)^T$$

such that

$$(y_1\mathbf{1}_m, y_2\mathbf{1}_m, \ldots, y_s\mathbf{1}_m)Q = (y_1\mathbf{1}_m, y_2\mathbf{1}_m, \ldots, y_s\mathbf{1}_m).$$

and hence 1 is an eigenvalue of Q.

Next we note that all the eigenvalues of Q are less than or equal to one [4]. Since the spectral radius of Q is always less than or equal to any matrix norm of Q and Q is irreducible, there is exactly one eigenvalue of Q equal to one. This implies that

$$\lim_{n\to\infty} Q^n = \mathbf{v}\mathbf{u}^T,$$

for certain non-zero vector $\mathbf{u}$ and $\mathbf{v}$. Therefore

$$\lim_{n\to\infty} \mathbf{x}_{n+1} = \lim_{n\to\infty} Q\mathbf{x}_n = \lim_{n\to\infty} Q^n\mathbf{x}_0 = \mathbf{v}\mathbf{u}^T\mathbf{x}_0 = \alpha\mathbf{v}.$$

Here α is a positive number since $\mathbf{x} \neq 0$ and is nonnegative. This implies that $\mathbf{x}_n$ tends to a stationary vector as n goes to infinity. Finally, we note that if $\mathbf{x}_0$ is a vector such that

$$\sum_{i=1}^{m} [\mathbf{x}_0^{(j)}]_i = 1, \ 1 \le j \le s,$$

then $Q\mathbf{x}_0$ and $\mathbf{x}$ are also vectors having this property. Hence the result follows. □

We remark that in the above proposition we only require a mild condition that $[\lambda_{ij}]$ is irreducible. While in [4], the authors assume that λ_{ij} are all positive. We note that $\mathbf{x}$ is not a probability distribution vector, but $\mathbf{x}^{(j)}$ is a probability distribution vector. The above proposition suggests one possible way to estimate the model parameters λ_{jk}. The idea is to find λ_{jk} which minimizes $||Q\hat{\mathbf{x}} - \hat{\mathbf{x}}||$ under certain vector norm $||\cdot||$.

3 Estimations of Model Parameters

In this section, we propose simple methods for the estimations of $P^{(jj)}$ and λ_{jk}. For each data sequence, one can estimate the transition probability matrix by the following method [4, 5, 6]. Given a data sequence, one can get the transition frequencies from one arbitrary state to the other states. Hence we can construct the transition frequency matrix for each of the data sequences. After making a normalization, the estimates of the transition probability matrices can also be obtained. We note that one has to estimate $O(s \times m^2)$ elements in transition frequency matrix for the multivariate Markov chain model. The vector $\mathbf{x}$ can be estimated from proportion of the occurrence of each state in each of the sequences. According to the idea at the end of last section, if we take $||.||$ to be $||.||_\infty$ we can get the values of λ_{jk} by solving the following optimization problem ([4, 5, 6]):

$$\begin{cases} \min\limits_{\lambda} \max\limits_{i} \left| \left[\lambda_{jj} \hat{P}^{(jj)} \hat{\mathbf{x}}^{(j)} + \sum\limits_{k=1, k\neq j}^{m} \lambda_{jk} \hat{\mathbf{x}}^{(k)} - \hat{\mathbf{x}}^{(j)} \right]_i \right| \\ \text{subject to} \\ \sum\limits_{k=1}^{s} \lambda_{jk} = 1, \quad \text{and} \quad \lambda_{jk} \geq 0, \quad \forall k. \end{cases} \tag{3.1}$$

Problem (3.1) can be formulated as s linear programming problems as follows, see for instance [8, (p. 221)]. For each j:

$$\min_{\lambda} w_j$$

subject to

$$\begin{cases} \begin{pmatrix} w_j \\ w_j \\ \vdots \\ w_j \end{pmatrix} \geq \hat{\mathbf{x}}^{(j)} - B_j \begin{pmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \vdots \\ \lambda_{js} \end{pmatrix}, \\ \begin{pmatrix} w_j \\ w_j \\ \vdots \\ w_j \end{pmatrix} \geq -\hat{\mathbf{x}}^{(j)} + B_j \begin{pmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \vdots \\ \lambda_{js} \end{pmatrix}, \\ w_j \geq 0, \\ \sum\limits_{k=1}^{s} \lambda_{jk} = 1, \quad \lambda_{jk} \geq 0, \quad \forall j, \end{cases} \tag{3.2}$$

where

$$B_j = [\hat{\mathbf{x}}^{(1)} \mid \hat{\mathbf{x}}^{(2)} \mid \cdots \mid \hat{P^{jj}} \hat{\mathbf{x}}^{(j)} \mid \cdots \mid \hat{\mathbf{x}}^{(s)}].$$

We remark that other vector norms such as $||.||_2$ and $||.||_1$ can also be used but they have different characteristics. The former will result in a quadratic programming problem while $||.||_1$ will still result in a linear programming problem, see for instance [8, (pp. 221-226)]. We note that the complexity of solving a linear programming problem or a quadratic programming problem is $O(n^3 L)$ where n is the number of variables and L is the number of binary bits needed to record all the data of the problem [9].

4 The Sales Demand Data Sequences

In this section, we present some numerical results based on the sales demand data of a soft-drink company in Hong Kong [4]. Products are categorized into six possible states according to sales volume. All products are labeled as either very fast-moving (very high sales volume), fast-moving, standard, slow-moving, very slow-moving (low sales volume) or no sales volume. The company has an important customer and would like to predict sales demand for this customer in order to minimize its inventory build-up and to maximize the demand satisfaction for this customer. Before developing a marketing strategy to deal with this customer, it is of great importance for the company to understand the sales pattern of this customer. Our multi-dimensional Markov chain model can be applied to model the sale demand data sequences and make predictions on the volume of the products in future based on the current and previous situations.

We first estimate all the transition probability matrices $P^{(jj)}$ by using the method proposed in Section 3 and we also have the estimates of the state distribution of the five products [4]. We use the multivariate Markov model to predict the next state $\hat{\mathbf{x}}_t$ at time t, which can be taken as the state with the maximum probability, i.e.,

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

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

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

where T is the length of the data sequence and

$$\delta_t = \begin{cases} 1, & \text{if } \hat{\mathbf{x}}_t = \mathbf{x}_t \\ 0, & \text{otherwise.} \end{cases}$$

Another way to compare the performance of the models is to use the BIC (Bayesian Information Criterion) [12] which is defined as

$$BIC = -2L + q \log n,$$

where

$$L = \sum_{j=1}^{s} \Big(\sum_{i_0,k_1,\cdots k_s=1}^{m} n^{(j)}_{i_0,k_1,\cdots,k_s} \log \Big(\sum_{l=1}^{m} \sum_{k=1}^{s} \lambda_{jk} p^{(jk)}_{i_0,k_l} \Big) \Big),$$

is the log-likelihood of the model,

$$n^{(j)}_{i_0,k_1,k_2,\cdots,k_s} = \sum x^{(j)}_{n+1}(i_0) x^1_n(k_1) x^2_n(k_2) \cdots x^s_n(k_s)$$

Here q is the number of independent parameters, and n is the length of the sequence. The less the value of BIC, the better the model is.

For the sake of comparison, we give numerical results of our new simplified model and the model proposed by Ching et al. [4] in Table 1. Although the results are more or less competitive when compared to the model in [4], it involves less variables. In Table 2, we give the BIC of the models. One can see that the simplified multivariate Markov model is much better than the multivariate Markov model in [4] in fitting the sales demand data. We remark that when $||.||_1$ is used instead of $||.||_\infty$, in the LP, we still get the same results for the prediction accuracy and BIC.

Models	A	B	C	D	E
The Multivariate Markov Model in [4]	50%	45%	63%	52%	55%
The Simplified Model	46%	46%	63%	52%	54%

Table 1. The Prediction Accuracy.

Models	BIC
The Multivariate Markov Model in [4]	8.0215e+003
The Simplified Model	3.9878e+003

Table 2. The BIC.

5 A Simplified Model for Very Short Sequences

In this section, we consider the case when the length of the observed data sequences are very short. In this case, we have two problems:
(a) the estimation of the transition probability matrices may have large error; and
(b) the steady-state may not be reached.

For Problem (a) we propose to replace the transition probability matrix $P^{(ii)}$ in Model (2.3) by the following rank-one matrix

$$(\hat{\mathbf{x}}^{(i)})^T(1,1,\ldots,1). \tag{5.1}$$

For Problem (b), the weights λ_{ij} should be chosen such that the multivariate Markov process converges very fast to the stationary distributions. The convergence rate of the process depends on the second largest eigenvalue in modulus of the matrix Q in (2.3). The reason is that the evolution process of the multivariate Markov chain is equivalent to the iterations of the power method. From numerical experience, the second largest eigenvalue depends very much on the value of λ_{ii}. We modified our simplified model for very short sequences by adding the extra constraints

$$0 \le \lambda_{ii} \le \beta. \tag{5.2}$$

They serve the purpose of controlling the second largest eigenvalue of Q and hence the convergence rate of the multivariate Markov chain.

Here we give an analysis of the simplified model with the assumptions (5.1) and (5.2) by further assuming that

$$P = P^{(ii)} = (\hat{\mathbf{x}})^T(1,1,\ldots,1) \quad \text{for all } i = 1,2,\ldots,s.$$

In this case, for $\lambda_{ij} > 0$ the steady-state probability distributions $\hat{\mathbf{x}}$ is an invariant. The problem here is how to assign λ_{ij} such that the second largest eigenvalue of Q is small. For simplicity of discussion, we assume one possible form for $[\Lambda]$ as follows:

$$\lambda_{ij} = \begin{cases} \lambda & \text{if} \quad i = j \\ \frac{1-\lambda}{m-1} & \text{if} \quad i \neq j \end{cases}$$

where $0 < \lambda < 1$. With these assumptions, we have the tensor product form for

$$Q = I \otimes \lambda P + (\Lambda - \lambda I) \otimes I.$$

Since the eigenvalues of P are given by 1 and 0 and the eigenvalues of Λ are given by

$$1 \quad \text{and} \quad \lambda - \frac{1-\lambda}{m-1}.$$

Here 1 is a simple eigenvalue in both cases. The eigenvalues of Q are then given by

$$1, \quad 1-\lambda, \quad \frac{\lambda-1}{m-1}, \quad \text{and} \quad \frac{\lambda m-1}{m-1}$$

where 1 and $1-\lambda$ are the two simple eigenvalues. The second largest eigenvalue of Q can be minimized by solving the following maxmin problem:

$$\min_{0<\lambda<1}\left\{\max\left\{1-\lambda, \frac{\lambda-1}{m-1}, \frac{\lambda m-1}{m-1}\right\}\right\}.$$

It is straight forward to check that the optimal value is

$$\lambda^* = \frac{m}{2m-1}$$

and the optimal second largest eigenvalue in this case is

$$\frac{m}{2m-1}.$$

6 Concluding Remarks

In this paper, we proposed a simplified multivariate Markov chain model for modelling categorical date sequences. The model is then applied to demand predictions. We also proposed a simplified model for the case when the observed data sequences are too short so that the steady-state may not be reached and the estimations of the transition probability matrices may not be accurate. The followings are some possible extensions of our models.

(i) Our multivariate Markov chain model is of first-order, one may further generalize the model to a higher-order multivariate Markov model, see for instance [7, 11]. We expect better prediction of sales demand can be achieved by using higher-order model.

(ii) Further research can also be done in extending the model to handle the case of "negative correlations". In the proposed models here, all the parameters λ_{ij} are assumed to be non-negative, i.e. the sequences j is "positively correlated" to the sequence i. It is interesting to allow λ_{ij} to be free.

Acknowledgment

The authors would like to thank the two anonymous referees and Prof. Gong-Yun Zhao for their helpful comments.

References

[1] P. Avery, The analysis of intron data and their use in the detection of short signals, *J. Mol. Evoln.* 26 (1987) 335–340.

[2] F. Blattner, G. Plunkett, C. Boch, N. Perna, V. Burland, M. Riley, J. Collado-Vides, J. Glasner C. Rode, G. Mayhew, J. Gregor, N. Davis, H. Kirkpatrick, M. Goeden, D. Rose, B. Mau, Y. Shao, *The complete genome sequence of Escherichia coli K-12*, 227 (1997) 1453–1462.

[3] W. Ching, Markov modulated Poisson processes for multi-location inventory problems, *Inter. J. Prod. Econ.* 53 (1997) 232–239.

[4] W. Ching, E. Fung and M. Ng, A multivariate Markov chain model for categorical data sequences and its applications in demand prediction, *IMA J. Manag. Math.* 13 (2002) 187–199.

[5] W. Ching, E. Fung and M. Ng, A higher-order Markov model for the Newsboy's problem, *J. Operat. Res. Soc.* 54 (2003) 291–298.

[6] W. Ching, E. Fung, M. Ng and T. Akutsu, On construction of stochastic genetic networks based on gene expression sequences, *Inter. J. Neural Systems* 15 (2005) 297–310.

[7] W. Ching, E. Fung and M. Ng, Higher-order Markov chain models for categorical data sequences, *Nav. Res. Logist.* 51 (2004) 557–574.

[8] V. Chvátal, *Linear Programming*, Freeman, New York, 1983.

[9] S. Fang and S. Pthenpura, *Linear Optimization and Extension*, Prentice-Hall, London, 1993.

[10] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, U.K, 1985.

[11] I. MacDonald and W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series*, Chapman & Hall, London, 1997.

[12] A. Raftery, A odel of high-order Markov chains, *J. Royal Statist. Soc.* 47 (1985) 528–539.

[13] A. Raftery and S. Tavare, Estimation and modelling repeated patterns in high-order Markov chains with the mixture transition distribution model, *Appl. Statist.* 43 (1994) 179–199.

[14] T. Siu, W. Ching, M. Ng and E. Fung, On multivariate credibility approach for portfolio credit risk measurement, *Quan. Fin.* 5 (2005) 543–556.

Manuscript received 15 November 2005
revised 15 June 2006
accepted for publication 15 June 2006

Wai-Ki Ching
Advanced Modeling and Applied Computing Laboratory, Department of Mathematics
The University of Hong Kong, Pokfulam Road, Hong Kong
E-mail address: wching@hkusua.hku.hk

Shu-Qin Zhang
Advanced Modeling and Applied Computing Laboratory, Department of Mathematics
The University of Hong Kong, Pokfulam Road, Hong Kong
E-mail address: sqzhang@hkusua.hku.hk

Michael K. Ng
Department of Mathematics, Hong Kong Baptist University
Kowloon Tong, Kowloon, Hong Kong
E-mail address: mng@math.hkbu.edu.hk