UDC 519.22

Statistical analysis of big data based on parsimonious models of high-order Markov chains

Yu. S. Kharin*

* *Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, Independence av. 4, Minsk, 220030, Belarus*

Abstract. The paper is devoted to construction of parsimonious (small-parametric) models of high-order Markov chains and to statistical inferences on parameters of these models.

Keywords: high-order Markov chain, parsimonious model, statistical analysis, big data.

1. Introduction

Applications in genetics, finance, medicine, information protection and other fields need to develop theory of statistical modeling and analysis of big data presented in the form of long discrete time series. An universal long-memory model for such data is the homogeneous Markov chain of sufficiently large order s on some finite state space A, $|A| = N$, $2 \le N < +\infty$. Unfortunately, the payment for this universality is exponential w.r.t. the order s number of parameters $D = O\left(N^{s+1}\right)$. To identify such a model we need to have "big data" sets and the computational complexity of order $O\left(N^{s+1}\right)$. To avoid this "curse of dimensionality" we propose to use the so-called parsimonious ("small-parametric") models of high-order Markov chains that are determined by small number of parameters. This paper presents probabilistic properties and statistical inferences on known and new parsimonious models.

2. Parsimonious models of high-order Markov chains and statistical inferences

Let $x_t \in A$ be a homogeneous Markov chain of the order s on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ determined by the $(s+1)$-dimensional matrix $P = (p_{i_1,\dots,i_s,i_{s+1}})$ of 1-step transition probabilities: $\mathbf{P}\{x_t = i_t | x_{t-1} = i_{t-1}, \dots, x_{t-s} = i_{t-s}\} = p_{i_{t-s},\dots,i_{t-1},i_t}$, $t > s$, $i_{t-s}, \dots, i_t \in A$.

The Jacobs – Lewis model with $D_{\mathrm{JL}} = N + s - 1$ parameters $\pi = (\pi_k)$, $\lambda = (\lambda_i)$, $\rho \in [0, 1]$ is determined by the equation:

$$x_t = \mu_t x_{t-\eta_t} + (1 - \mu_t)\xi_t, \tag{1}$$

where $t > s$, $\{\xi_t, \eta_t, \mu_t\}$ are jointly independent random variables:

$$\mathbf{P}\{\mu_t = 1\} = 1 - \mathbf{P}\{\mu_t = 0\} = \rho;$$

$$\mathbf{P}\{\eta_t = i\} = \lambda_i,\ i \in \{1, 2, \ldots, s\},\ \sum_{i=1}^{s} \lambda_i = 1,\ \lambda_s \neq 0;$$

$$\mathbf{P}\{\xi_t = k\} = \pi_k,\ k \in A,\ \sum_{k \in A} \pi_k = 1; \tag{2}$$

$$\mathbf{P}\{x_1 = k\} = \ldots = \mathbf{P}\{x_s = k\} = \pi_k,\ k \in A.$$

Theorem 1. *The discrete-valued time series x_t determined by* (2), (2) *is a homogeneous Markov chain of the order s with the initial probability distribution $\pi_{i_1,\ldots,i_s} = \pi_{i_1} \cdot \ldots \cdot \pi_{i_s}$ and the $(s+1)$-dimensional matrix of transition probabilities $P(\pi, \lambda, \rho) = (p_{i_1,\ldots,i_{s+1}})$:*

$$p_{i_1,\ldots,i_s,i_{s+1}} = (1-\rho)\pi_{i_{s+1}} + \rho \sum_{j=1}^{s} \lambda_j \mathrm{I}\{i_{s-j+1} = i_{s+1}\},\ i_1, \ldots, i_{s+1} \in A,$$

where $\mathrm{I}\{\cdot\}$ *is the indicator function.*

By Theorem 1 we construct some consistent estimators by the observations $X_n = (x_1, \ldots, x_n)$ that are used as initial values in the iterative computation of the MLEs $\hat{\pi}$, $\hat{\lambda}$, $\hat{\rho}$; we also construct statistical tests for true values of parameters [1].

The Raftery MTD (Mixture Transition Distribution)-model:

$$p_{i_1,\ldots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j q_{i_j,i_{s+1}},\ i_1, \ldots, i_{s+1} \in A, \tag{3}$$

where $Q = (q_{i,k})$ is a stochastic $(N \times N)$-matrix, $\lambda = (\lambda_1, \ldots, \lambda_s)'$ is a discrete probability distribution, $\lambda_1 > 0$.

Theorem 2. *For the ergodic MTD-model* (2) *the 2-dimensional stationary probability distribution of the random vector $(x_{t-m}, x_t)'$, $1 \leq m \leq s$, has the form:*

$$\pi^*_{ki}(m) = \pi^*_k \pi^*_i + \pi^*_k \lambda_{s-m+1}(q_{ki} - \pi^*_i),\ i, k \in A.$$

Using Theorem 2 we construct consistent asymptotically unbiased estimators $\tilde{Q}$, $\tilde{\lambda}$ [1].

Markov chain of the order s with r partial connections:

$$p_{J_1^{s+1}} = p_{j_1,\ldots,j_s,j_{s+1}} = q_{j_{m_1^0},\ldots,j_{m_r^0},j_{s+1}},\ J_1^{s+1} \in A^{s+1}, \tag{4}$$

where r is the number of connections; $M_r^0 = (m_1^0, \ldots, m_r^0)$ is the integer valued vector with r ordered components $1 = m_1^0 < m_2^0 < \ldots < m_r^0 \leq s$, called

the connection template; $Q = \left(q_{J_1^{r+1}}\right)_{J_1^{r+1} \in A^{r+1}}$ is an $(r+1)$-dimensional stochastic matrix. If $r = s$ we have the full-connected Markov chain of the order s.

For the model (2) we construct consistent statistical estimators for the parameters $\hat{Q}$, $\hat{M}_r$, $\hat{r}$, $\hat{s}$, statistical tests for the true values of Q; the performance characteristics of these statistical inferences are also given [2].

Introduce the notation: $1 \le L \le s-1$, $K = N^L - 1$, $1 \le M \le K+1$ are some positive integers; $Q^{(1)}, \ldots, Q^{(M)}$ are M different quadratic stochastic matrices of the order N: $Q^{(m)} = (q_{ij}^{(m)})$; $< J_n^m >= \sum\limits_{k=n}^{m} N^{k-n} j_k$ is the numeric representation of the multiindex $J_n^m = (j_n, j_{n+1}, \ldots, j_m) \in A^{m-n+1}$.

Markov chain of conditional order:

$$p_{J_1^{s+1}} = \sum_{k=0}^{K} \mathrm{I}\{< J_{s-L+1}^{s} >= k\} q_{j_{b_k}, j_{s+1}}^{(m_k)}, \; J_1^{s+1} \in A^{s+1}, \tag{5}$$

where $1 \le m_k \le M$, $1 \le b_k \le s-L$, $0 \le k \le K$, $\min\limits_{0 \le k \le K} b_k = 1$, the sequence J_{s-L+1}^{s} is called the base memory fragment of the random sequence. Number of parameters: $D_{\mathrm{MCCO}} = 2(N^L + 1) + MN(N-1)$.

Statistical inferences for this new model (2) are considered in [3].

One more new parsimonious model is the binary conditionally nonlinear autoregressive model.

Theoretical results for all these models are illustrated by results of computer experiments on simulated and real data.

References

1. *Kharin Yu.* Robustness in Statistical Forecasting. — Springer: N.Y., 2013.
2. *Kharin Yu, Piatlitski A.* Markov chain of order s with r partial connections and statistical inference on its parameters // Discrete Mathematics and Applications. — 2007. — Vol. 17, no. 3. — P. 295–317.
3. *Kharin Yu., Maltsew M.* Algorithms for statistical analysis of Markov chain with conditional memory depth. // Informatics. — 2011. — No. 1. — P. 34–43.