DISCRETE-VALUED TIME SERIES IN COMPUTER DATA ANALYSIS AND INFORMATION SECURITY

Yu. S. Kharin

Belarusian State University
Minsk, Republic of Belarus
E-mail: Kharin@bsu.by

An analytical review of models for discrete-valued time series based on high-order Markov chains is presented. A new model – Markov chain of the order s with r partial connections $\mathbf{MC}(s,r)$ − is developed. Computer algorithms of statistical analysis for this model are given and analyzed theoretically and on real statistical data. The results are applicable in computer data analysis and information security.

Key words: discrete-valued time series, high-order Markov chains, estimation.

INTRODUCTION

Discrete-valued time series are required for mathematical and computer modeling of complex systems and IT in many applied fields [1–12]: in bioinformatics for recognition and analysis of genetic sequences, in medical diagnostics, in economics and finance to forecast the dynamics of economic and financial indicators, in meteorology to forecast the weather, in sociology for modeling of social behavior, in Internet-traffic analysis for optimizing protocols, in computer networks for evolution of information security.

Discrete-valued time series is a random process $x_t \in \mathbf{A}$ on some probability space $(\mathbf{\Omega}, \mathbf{F}, \mathbf{P})$ with discrete time $t \in \mathbf{N}_0 = \{0,1,2,\ldots\}$ and a discrete state space $\mathbf{A}$ with the cardinality $|\mathbf{A}| = N$, $2 \le N \le +\infty$. If $\mathbf{A} = \mathbf{N}_0$ is the countable state space ($N = +\infty$), in the literature the models are based on the so-called Integer Autoregression Model of order $p \in \boldsymbol{N}_0$ ($INAR(p)$) determined by a stochastic difference equation [1]

$$x_t = \sum_{j=1}^{x_{t-1}} \xi_{t1j} + \ldots + \sum_{j=1}^{x_{t-p}} \xi_{tpj} + \eta_t, \quad t \in \mathbf{N}_0,$$

where for all $t \in \mathbf{N}_0$ and $i \in \{1,\ldots,p\}$, $\{\xi_{tij} : j \in \mathbf{N}_0\}$ is a sequence of independent identically distributed Bernoulli $Bi(1, \alpha_i)$ random variables, $\{\varepsilon_t\}$ is a sequence of independent random variables, and $\{\xi_{tij}\}$, $\{\varepsilon_t\}$, $\{x_0,\ldots,x_{-p+1}\}$ are independent.

If $\mathbf{A}$ is a finite state space (without loss of generality $\mathbf{A} = \{0,1,\ldots,N-1\}$, $2 \le N < +\infty$), the time series x_t in this case is called in some papers as the categorical time series [4]; a review of results is also given in [4]. To treat this case, we propose in the paper the small-parametric approach based on high-order Markov chains.

LONG-MEMORY DISCRETE-VALUED TIME SERIES

For discrete-valued time series, as for «continious» time series (where $\mathbf{A} = (-\infty, +\infty)$ is the set of real numbers), much attention is paid now to «long-memory» models. An universal «long-memory» model for discrete-valued time series x_t is the homogeneous Markov chain $\mathbf{MC}(s)$ of the order $s \in \mathbf{N}_0$, determined by the generalized Markov property ($t > s$): $\mathbf{P}\{x_t = i_t | x_{t-1} = i_{t-1}, \ldots, x_1 = i_1\} = \mathbf{P}\{x_t = i_t | x_{t-1} = i_{t-1}, \ldots, x_{t-s} = i_{t-s}\} = p_{i_{t-s},\ldots,i_{t-1},i_t}$,
where s is the memory length; $i_1, i_2, \ldots, i_t \in \mathbf{A}$ are values of the process at the time moments $1,2,\ldots,t$; $P = \left(p_{i_{t-s},\ldots,i_{t-1},i_t}\right)$ is an $(s+1)$-dimensional matrix of one-step transition probabilities. Number of independent parameters for the $\mathbf{MC}(s)$-model increases exponentially w.r.t. the memory length s: $D_{\mathbf{MC}(s)} = N^s(N-1)$.

To identify this model we need to have huge data sets and the computation work of size $\mathrm{O}\left(N^{s+1}\right)$. To avoid this «curse of dimensionality» we propose to use the «small-parametric» models of high-order Markov chains that are determined by small number of parameters $d \ll D_{\mathbf{MC}(s)}$.

Jacobs-Lewis model is determined by a stochastic difference equation [3] ($t > s$):

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

where $\{\xi_t, \eta_t, \mu_t\}$ are independent random variables with probability distributions:

$$\mathrm{P}\{\mu_t = 1\} = 1 - \mathrm{P}\{\mu_t = 0\} = \rho;\ \mathrm{P}\{\xi_t = k\} = \pi_k,\ k \in \mathbf{A},\ \sum_{k\in \mathrm{A}} \pi_k = 1; \qquad (2)$$

$$\mathrm{P}\{\eta_t = 1\} = \lambda_i,\ i \in \{1,2,\dots,s\},\ \sum_{i=1}^{s} \lambda_i = 1,\ \lambda_s \neq 0.$$

Number of parameters depends linearly on s: $D_{JL} = N + s - 1$.

In [3] only moments and stationary distributions were analyzed. We proved [6, 7] probabilistic and statistical properties of the model (1), (2) by $\mathbf{MC}(s)$-model.

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

$$p_{i_1,\dots,i_s,i_{s+1}} = (1-\rho)\pi_{i_{s+1}} + \sum_{j=1}^{s} \lambda_j \delta_{i_{s-j+1},i_{s+1}} \quad i_1,\dots,i_{s+1} \in \mathbf{A},$$

where $\delta_{j,k}$ is the Kronecker symbol.

Corollary 1. Maximum likelihood estimators (MLE) $(\hat{\pi},\hat{\lambda},\hat{\rho})$ by the data $X_1^n = (x_1,\dots,x_n)'$ are determined by the solution of the maximization problem:

$$l(\pi,\lambda,\rho) = \sum_{t=1}^{s} \ln \pi_{x_t} + \sum_{t=t+1}^{n} \ln\left((1-\rho)\pi_{x_t} + \rho\sum_{j=1}^{s} \lambda_j \delta_{x_{t-j},x_t}\right) \to \max_{\pi,\lambda,\rho}.$$

Using these MLE we had constructed [6, 7] the consistent generalized probability ratio test for hypotheses on the true values of parameters (ρ,λ,π) in (2).

Mixture Transition Distribution–model (MTD-model) was proposed in 1985 by A. Raftery [12] as a special small-parametric representation:

$$p_{i_1,\dots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j q_{i_j,i_{s+1}},\ i_1,\dots,i_{s+1} \in \mathbf{A}, \qquad (3)$$

where $Q = (q_{i,k})$ is a stochastic $(N\times N)$-matrix, $0 \le q_{i,k} \le 1$, $\sum_{k\in\mathbf{A}} q_{i,k} \equiv 1$, $i,k \in \mathbf{A}$, $\lambda = (\lambda_1,\dots\lambda_s)'$ is a discrete probability distribution, $\lambda_1 > 0$.

The MTDg (generalized MTD)-model:

$$p_{i_1,\dots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j q^{(j)}_{i_{s-j+1},i_{s+1}},\ i_1,\dots,i_{s+1} \in \mathbf{A}, \qquad (4)$$

where $Q^{(j)} = (q^{(j)}_{i,k})$ is a stochastic matrix for the j-th lag.

Number of parameters for the MTDg: $D_{MTDg} = s(N(N-1)/2+1) - 1$.

We have constructed a simple criterion for the ergodicity of the MTD-model and found a useful property of the stationary probability distribution [6].

Theorem 2. For the MTDg-model (4), if $\exists k \in N: \left(\left(Q^{(1)} \right)^K \right)_{ij} > 0$, $\forall i, j \in \mathbf{A}$, then the s-dimensional stationary p.d. satiafies the equation $(i_1, \dots, i_s \in \mathbf{A})$:

$$\pi^*_{i_1,\dots,i_s} = \prod_{l=0}^{s-1} \left(\pi^*_{i_s-l} + \sum_{j=l+1}^{s} \lambda_j \left(q^{(j)}_{i_{j-l},i_{s-l}} - \sum_{r=0}^{N-1} q^{(j)}_{r,i_{s-l}} \pi^*_r \right) \right).$$

Corollary 2. For the ergodic MTD-model (3) the 2-dimensional stationary p.d. of the random vector $(x_{t-m}, x_t)'$ $\pi^*_{ki}(m) = \pi^*_k \pi^*_i + \pi^*_k \lambda_{s-m+1} \left(q_{ki} - \pi^*_i \right)$, $i,k \in \mathbf{A}$, $1 \le m \le s$.

Based on Corollary 2 we have constructed statistical estimators for λ, Q by an observed time series $X_1^n = (x_1, \dots, x_n)'$ of the length n:

$$\tilde{\pi}_i = \frac{1}{n-2s+1} \sum_{t=s+1}^{n-s+1} \delta_{x_t,i}\ ;\ \tilde{\pi}_{ki}(j) = \frac{1}{n-2s+1} \sum_{t=s+j}^{n-s+j} \delta_{x_{t-j},k} \delta_{x_t,i}\ ;$$

$$z_{ki}(j) = \tilde{\pi}_{ki}(s-j) / \tilde{\pi}_k - \tilde{\pi}_i,\ d_{ki} = \tilde{q}_{ki} - \tilde{\pi}_i,\ i,k \in \mathbf{A};\ \tilde{\boldsymbol{\lambda}}_j = \sum_{i,k \in \mathbf{A}} z_{ki}(s-j) d_{ki} / \sum_{i,k \in \mathbf{A}} d_{ki}^2,\ j = 1,\dots,s;$$

$$\tilde{q}_{ki} = \left\{ \sum_{j=1}^{s} \tilde{\pi}_{ki}(j) \ /\tilde{\pi}_k - (s-1)\tilde{\pi}_i, \quad \text{if} \quad \tilde{\pi}_k > 0;\ N^{-1} \ \text{else} \right\}. \tag{5}$$

Theorem 3. For the ergodic MTD-model (3) the estimators $\tilde{Q}$, $\tilde{\lambda}$ determined by (5) at $n \to \infty$ are consistent and asymptotically unbiased.

MLE $\hat{Q}$, $\hat{\lambda}$ are solutions of the nonlinear maximization problem:

$$l(Q,\lambda) = \sum_{t=s+1}^{n} \ln \sum_{j=1}^{s} \lambda_j q_{x_{t-s+j-1},x_t} \to \max_{Q,\lambda}. \tag{6}$$

The estimators (5) $\tilde{Q}$, $\tilde{\lambda}$ are used as initial values in the iterative computation of the MLEs $\hat{Q}$, $\hat{\lambda}$ in (6). Generalized probability ratio test of the asymptotic size $\boldsymbol{\varepsilon} \in (0,1)$ for H_0: $Q = Q^0$, $\lambda = \lambda^0$, $H_1 = \bar{H}_0$ is constructed as in the previous section.

MARKOV CHAIN MC(s, r) OF THE ORDER s WITH r PARTIAL CONNECTIONS AND ITS PROPERTIES

The $\mathbf{MC}(s,r)$ proposed by Yu. Kharin in 2004 [5] is determined by the following small-parametric reparametrization of the $(s+1)$-dimensional transition probability matrix:

$$p_{J_1^{s+1}} = p_{j_1,\dots,j_s,j_{s+1}} = q_{j_{m_1^0},\dots,j_{m_r^0},j_{s+1}}, \tag{7}$$

where $J_1^{s+1} = (j_1, \dots, j_{s+1})$ is the $(s+1)$-dimensional index vector; r is the number of connections $(1 \le r \le s)$; $M_r^0 = (m_1^0, \dots, m_r^0) \in M$ is the integer-valued vector with r ordered components, $1 = m_1^0 < m_2^0 < \dots < m_r^0 \le s$, called the pattern of connections; $Q = (q_{J_1^{r+1}})_{J_1^{r+1} \in \mathbf{A}^{r+1}}$ is an $(r+1)$-dimensional stochastic matrix. If $r = s$, we have the general $\mathbf{MC}(s)$ model.

In [5, 9, 11] the probabilistic properties of the $\mathbf{MC}(s,r)$-model are found.

Theorem 4. The $\mathbf{MC}(s,r)$ defined by (7) is an ergodic Markov chain if and only if there exists $i \in N$ such that $\min_{J_1^s, J_{s+i+1}^{2s+i} \in \mathbf{A}^s} \sum_{J_{s+1}^{s+i} \in \mathbf{A}^i} \prod_{k=1}^{s+i} q_{j_{k+m_1^0-1},\dots,j_{k+m_r^0-1},j_{k+s}} > 0$. Stationary probability distribution $\left(\pi^*_{J_1^s}\right)_{J_1^s \in \mathbf{A}^s}$ satisfies the equations: $\boldsymbol{\pi}^*_{J_2^{s+1}} = \sum_{j_1 \in \mathbf{A}} \boldsymbol{\pi}^*_{J_1^s} q_{j_{m_1^0},\dots,j_{m_r^0},j_{s+1}}$, $J_1^{s+1} \in \mathbf{A}^s$.

Corollary 3. For a stationary Markov chain the stationary probability distribution has the multiplicative form $\pi^*_{J_2^s} = \prod_{i=1}^{s} \pi^*_{j_i}$, $J_1^s \in \mathbf{A}^s$, if and only if $\pi^*_{j_{r+1}} = \sum_{j_1 \in \mathrm{A}} \pi^*_{j_i} q_{J_1^{r+1}}$, $J_2^{r+1} \in \mathbf{A}^r$.

Corollary 4. If Q is doubly stochastic: $\sum_{j_1 \in \mathbf{A}} q_{J_1^{r+1}} \equiv 1$, $\sum_{j_{r+1} \in \mathbf{A}} q_{J_1^{r+1}} \equiv 1$, then the stationary probability distribution is uniform: $\pi^*_{J_1^s} \equiv N^{-s}$.

STATISTICAL ANALYSIS BASED ON THE MC(*s*, *r*)-MODEL

Introduce the notation: $F\left(J_i^{i+s-1}; M_r\right) = \left(j_{i+m_1-1}, \dots, j_{i+m_r-1}\right)$ is the selector-function;

$$\nu_{J_1^{r+1}}\left(X_1^n; M_r\right) = \sum_{t=1}^{n-s} \delta_{F\left(X_t^{t+s-1}; M_r\right), J_1^r} \delta_{x_{t+s}, j_{r+1}} \tag{8}$$

is the frequency statistic for a pattern $M_r \in M$; $\mu_{J_1^{r+1}}\left(M_r\right) = P\left\{F\left(X_t^{t+s-1}; M_r\right) = J_1^r, x_{t+s} = j_{r+1}\right\}$ is the probability distribution of the $(r+1)$-tuple; the dot used instead of any index means summation on all its values: $\mu_{J_1^r\cdot}\left(M_r\right) = \sum_{j_{r+1} \in \mathbf{A}} \mu_{J_1^{r+1}}\left(M_r\right)$; $\hat{\mu}_{J_1^{r+1}}\left(M_r\right) = \nu_{J_1^{r+1}}\left(X_1^n; M_r\right)/(n-s)$ is the frequency estimator for the probability $\mu_{J_1^{r+1}}\left(M_r\right)$, $J_1^{r+1} \in \mathbf{A}^{r+1}$, $M_r \in M$.

Theorem 5 [5]. If the pattern of connections M_r^0 is know n, then the MLE for the matrix Q is $\hat{Q} = \left(\hat{q}_{J_1^{r+1}}\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$: $\hat{q}_{J_1^{r+1}} = \left\{\hat{\mu}_{J_1^{r+1}}\left(M_r^o\right) / \hat{\mu}_{J_1^r\cdot}\left(M_r^o\right), \text{if } \hat{\mu}_{J_1^r\cdot}\left(M_r^o\right) > 0;\ 1/N \text{ else}\right\}$. Under the stationarity condition $\left\{\hat{q}_{J_1^{r+1}}:\ J_1^{r+1} \in \mathbf{A}^{r+1}\right\}$ are asymptotically ($n \to \infty$) unbiased and consistent with covariances $\mathbf{Cov}\left\{\hat{q}_{J_1^{r+1}}, \hat{q}_{K_1^{r+1}}\right\} = \sigma^{\hat{q}}_{J_1^{r+1}, K_1^{r+1}} / (n-s) + \mathrm{O}(1/n^2)$,

$$\sigma^{\hat{q}}_{J_1^{r+1}, K_1^{r+1}} = \delta_{J_1^r, K_1^r}\, \delta_{J_1^r, K_1^r}\, q_{J_1^{r+1}} \left(\delta_{j_{r+1}, k_{r+1}} - q_{K_1^{r+1}}\right) / \mu_{J_1^r\cdot}\left(M_r^0\right),\ J_1^{r+1}, K_1^{r+1} \in \mathbf{A}^{r+1}.$$

Moreover, the probability distribution of the N^{r+1}-dimensional random vector $\left(\sqrt{n-s}\left(\hat{q}_{J_1^{r+1}} - q_{J_1^{r+1}}\right)\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$ at $n \to \infty$ converges to the normal probability distribution with zero mean and the covariance matrix $\Sigma^{\hat{q}} = \left(\sigma^{\hat{q}}_{J_1^{r+1}, K_1^{r+1}}\right)$.

The consistent statistical test for the hypotheses H_0: $Q = Q^0$, where $Q^0 = \left(q^0_{J_1^{r+1}}\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$; $H_1 = \overline{H}_0$, consists of the following steps.

1. Computation of the statistics $\nu_{J_1^{r+1}}\left(X_1^n;M_r^0\right)$, $J_1^{r+1}\in \mathbf{A}^{r+1}$, by (8).
2. Computation of the statistic $\left(D_{J_1^r}=\left\{j_{r+1}\in \mathbf{A}: q^0_{J_1^{r+1}}>0\right\}\right)$

$$\rho=\sum_{J_1^r\in\mathbf{A}^r,\, j_{r+1}\in D_{J_1^r}} \nu_{J_1^r\cdot}\left(X_1^n;M_r^0\right)\left(\hat{q}_{J_1^{r+1}}-q^0_{J_1^{r+1}}\right)^2/q^0_{J_1^{r+1}}.$$

3. Computation of the P-value: $P=1-G_U(\rho)$, where $G_U(\cdot)$ is distribution the standard χ^2- distribution function with $U=\sum_{J_1^r\in\mathbf{A}^r}\left(\left|D_{J_1^r}\right|-1\right)$ degrees of freedom.
4. The decision rule (ε– asymp. signif. level): if $P\ge\varepsilon$, then to conclude that the hypothesis H_0 is true; otherwise, the alternative H_1 is true.

Corollary 5. Under stationary $\mathbf{MC}(s,r)$ and contigual family of alternatives H_{1n}: $Q=Q^{1n}$, where $Q^{1n}=\left(q^{1n}_{J_1^{r+1}}\right)_{J_1^{r+1}\in\mathbf{A}^{r+1}}$, $q^{1n}_{J_1^{r+1}}=q^0_{J_1^{r+1}}\dfrac{1+d_{J_1^{r+1}}}{\sqrt{n-s}}$, $\sum_{j_{r+1}\in\mathbf{A}} d_{J_1^{r+1}}q^0_{J_1^{r+1}}=0$, $\sum_{J_1^{r+1}\in\mathbf{A}^{r+1}}\left|d_{J_1^{r+1}}\right|>0$, if H_{1n} is true, then at $n\to\infty$ the power of the developed test, $w\to 1-G_{U,a}\left(G_U^{-1}(1-\varepsilon)\right)$, where $G_{U,a}(\cdot)$ is the probability distribution function of the noncentral χ^2 distribution with U degrees of freedom and the noncentrality parameter $a=\sum_{J_1^{r+1}\in\mathbf{A}^{r+1}}\mu_{J_1^{r+1}}\left(M_r^0\right)d^2_{J_1^{r+1}}$.

Introduce the notation: M is the set of all admissible patterns M_r;

$$H\left(M_r\right)=-\sum_{J_1^{r+1}\in\mathbf{A}^{r+1}}\mu_{J_1^{r+1}}\left(M_r\right)\ln\left(\mu_{J_1^{r+1}}\left(M_r\right)/\mu_{J_1^{r+1}\cdot}\left(M_r\right)\right)\ge 0 \tag{9}$$

is the conditional entropy of the future symbol $x_{t+s}\in\mathbf{A}$ relative to the past derived by the selector $F\left(X_t^{t+s-1};M_r\right)\in\mathbf{A}^r$, $M_r\in M$; $\widehat{H}\left(M_r\right)$ is the “plug-in” estimator of the conditional entropy, which is generated by substitution of true probabilities $\mu_{J_1^{r+1}}\left(M_r\right)$ in (9) by their estimators $\hat{\mu}_{J_1^{r+1}}\left(M_r\right)$.

Theorem 6. If the order s and the number of connections r are known, then the MLE $\widehat{M}_r=\arg\min_{M_r\in M}\widehat{H}\left(M_r\right)$. Under the stationarity condition of the $\mathbf{MC}(s,r)$ the estimator $\widehat{M}_r$ at $n\to\infty$ is consistent: $\widehat{M}_r\xrightarrow{\mathbf{P}}M_r^0$.

Let $s\in[s_-,s_+]$, $r\in[r_-,r_+]$, $1\le s_-<s_+<\infty$, $1\le r_-<r_+<s_+$. To estimate the parameters r, s we use the Bayesian Information Criterion (BIC), which has the form [9–11]:

$$BIC(s,r)=2(n-s)\widehat{H}\left(\widehat{M}_r\right)+U\ln(n-s), \tag{10}$$

where $U=\sum_{J_1^r\in\mathbf{A}^r}\left(\left|D_{J_1^r}\right|-1+\delta_{\hat{\mu}_{J_1^r\cdot}\left(\widehat{M}_r\right),0}\right)$, $D_{J_1^r}=\left\{j_{r+1}\in\mathbf{A}:\hat{\mu}_{J_1^{r+1}}\left(\widehat{M}_r\right)>0\right\}$

Consistent estimators $\hat{s}$, $\hat{r}$ are determined by minimization: $BIC(s,r)\to\min_{s_-\le s\le s_+,\, r_-\le r\le r_+}$.

APPLICATION OF THE MC(*s*, *r*) TO REAL STATISTICAL DATA

The developed algorithms were successfully tested on simulated data and also applied to real statistical data [6, 7, 11]. We present here only two examples: in meteorology and in genetics.

Modeling of wind direction. The discrete-valued time series of the daily average wind speed at Malin Head (North of Ireland) during the period 1961–1978 [12] $x_t \in \{0,1,2\}$ of the length $n = 6574$ was fitted by the $\mathbf{MC}(s,r)$-model for $s = \{1,2,...,7\}$, $r = \{1,2,...,7\}$. Table presents the values of the BIC for the different pairs (s,r).

Different modelings of the wind speed data

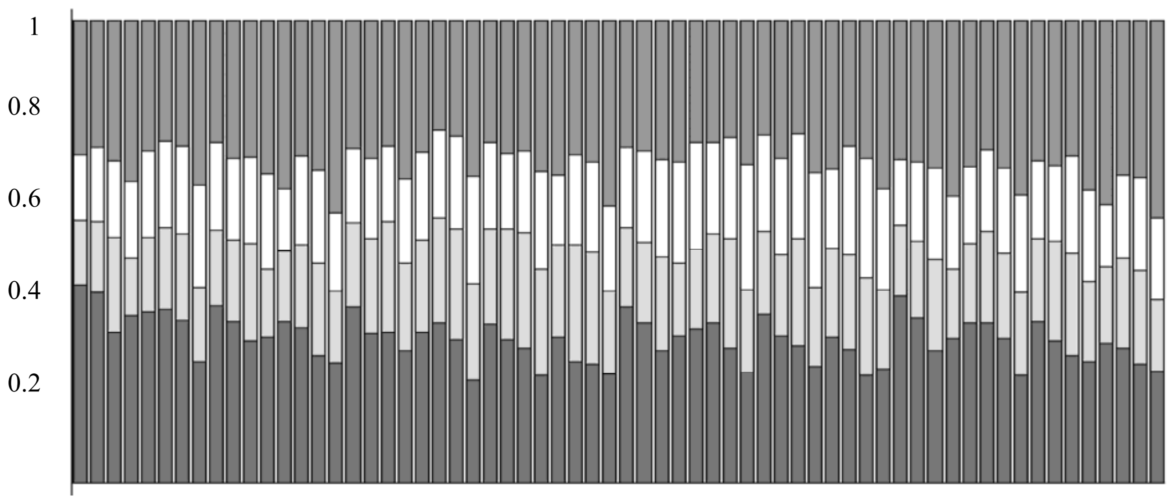
Model	BIC	Model	BIC	Model	BIC	Model	BIC
MC(1,1)	8127.52	MC(4,2)	8139.12	MC(5,5)	8621.97	MC(7,1)	9041.43
MC(2,1)	8777.63	MC(4,3)	8164.79	MC(6,1)	9016.23	MC(7,2)	8163.07
MC(2,2)	8096.08	MC(4,4)	8332.77	MC(6,2)	8148.48	MC(7,3)	8197.91
MC(3,1)	8849.90	MC(5,1)	8984.10	MC(6,3)	8190.78	MC(7,4)	8323.19
MC(3,2)	**8079.81**	MC(5,2)	8129.83	MC(6,4)	8350.82	MC(7,5)	8599.09
MC(3,3)	8143.13	MC(5,3)	8177.92	MC(6,5)	8576.92	MC(7,6)	8973.15
MC(4,1)	8956.11	MC(5,4)	8349.62	MC(6,6)	8969.54	MC(7,7)	9575.64

The best fitted model is the $\mathbf{MC}(3,2)$ with the pattern $\widehat{M}_r = (1,3)$ and the matrix

$$\widehat{Q}' = \begin{pmatrix} 0.27 & 0.08 & 0 & 0.22 & 0.04 & 0 & 0.21 & 0.02 & 0 \\ 0.73 & 0.86 & 0.63 & 0.78 & 0.82 & 0.52 & 0.79 & 0.72 & 0.43 \\ 0 & 0.06 & 0.37 & 0 & 0.14 & 0.48 & 0 & 0.26 & 0.57 \end{pmatrix}.$$

The fitted model $\mathbf{MC}(3,2)$ detects significant dependencies in the data.

Genomic sequencing for the drosophila genome sequence (www.fruitfly.org) $N = 4$, $n = 5 \cdot 10^5$. If $s_- = 1$, $s_+ = 8$, $r_- = 1$, $r_+ = 8$, then the best fitted model is the $\mathbf{MC}(6,3)$ with the pattern $\widehat{M}_r = (1,5,6)$ and the matrix $\widehat{Q}$ visualized in Figyre. Here on «x-axes» the values of $\widehat{M}_r$–prehistory are indicated, «y-axes» gives the values of one-step transition probabilities to four states indicated by different levels of grey colors.



The matrix $\widehat{Q}$ for the genomic sequencing

The $\mathbf{MC}(s,r)$ can be also useful in performance evaluation for generators of pseudo-random sequences [10].

REFERENCES

1. *Alzaid, A. A.* An integer-valued pth-order autoregressive structure (INAR(p)) process / A. A. Alzaid, M. Al-Osch // Journal of Applied Probability. 1990. Vol. 27. P. 314–324.

2. *Buhlmann, P.* Variable length Markov chains / P. Buhlmann, A. J. Wyner // The Annals of Statistics. 1999. Vol. 27, № 2. P. 480–513.

3. *Jacobs, P. A.* Discrete time series generated by mixtures I: correlational and runs properties / P. A. Jacobs, P. A. W. Lewis // Journal of the Royal Statistical Society. Ser. B.1978. Vol. 40, № 1. P. 94–105.

4. *Kedem, B.* Regression Models for Time Series Analysis / B. Kedem, K. Fokianos // John Wiley. 2002, 360 p.

5. *Kharin, Yu.* S. Markov chains with r-partial connections and their statistical estimators / Yu. S. Kharin // Doklady of NAN Belarusi. 2004. Vol. 48, № 1. P. 40–44 (in Russian).

6. *Kharin, Yu. S.* Optimality and Robustness in Statistical Forecasting. / Yu. S. Kharin // Minsk: BSU, 2008, 263 p. (in Russian).

7. *Kharin, Yu.* S. Probabilistic and statistical analysis of high-order Markov chains / Yu. S. Kharin // Vestnik BSU. 2006. Ser. 1, № 3. P. 80–86 (in Russian).

8. *Kharin, Yu.* S. Discriminant analysis of stationary finite Markov chains / Yu. S. Kharin, A. L. Kostevich // Mathematical Methods of Statistics. 2004. № 1. P. 235–252.

9. *Kharin, Yu. S.* A Markov chain of order s with r partial connections and statistical inference on its parameters / Yu. S. Kharin, A. I. Piatlitski // Discrete Mathematics and Applications. 2007. Vol. 17, № 3. P. 295–317.

10. *Kharin, Yu.* S. An approach to statistical testing of random and pseudorandom sequences / Yu. S. Kharin, A. I. Piatlitski // Management in Information Security. 2008. Vol. 12, № 3. P. 302–304 (in Russian).

11. *Kharin, Yu.* S. Statistical analysis of high-order Markov chains with partial connections / Yu. S. Kharin, A. I. Petlitski // Computer data analysis and Modeling. Vol. 1. P. 66–73.

12. *Raftery, A.* A model for high-order Markov chains / A. Raftery // Journal of the Royal Statistical Society. Ser. B. 1985. Vol. 47, № 3. P. 528–539.