Austrian Journal of Statistics
June 2014, Volume 43/3-4, 205–216.
http://www.ajs.or.at/

Markov Chain of Conditional Order: Properties and Statistical Analysis

Yuriy Kharin
Belarusian State University

Mikhail Maltsau
Belarusian State University

Abstract

The paper deals with finite Markov chain of conditional order, that is a special case of high-order Markov chain with a small number of parameters. Statistical estimators for parameters and statistical tests for parametric hypotheses are constructed and their properties are analyzed. Results of computer experiments on simulated and real data are presented.

Keywords: markov chain, conditional order, ergodicity, statistical estimator, hypothesis testing.

1. Introduction

Finite Markov chain of the order s $(1 \leq s < \infty)$ described by Doob (1953) is a well-known universal mathematical model to analyze long memory discrete-valued time series in many applied fields. It is used for statistical data analysis in genetics (see Waterman 1999), economics (see Ching 2004), signal processing (see Li, Dong, Zhang, Zhao, Shi, and Zhao 2010) and other areas.

Unfortunately, there is a significant disadvantage of this model. It has exponential complexity since the number of independent parameters $D(s)$ of the N-state Markov chain of the order s increases exponentially w.r.t. s:

$$D(s) = (N-1)N^s = O(N^{s+1}).$$

Because of the "curse of dimensionality" to identify this model one needs time series of big size (length of time series) $n \geq D(s)$ not available in practice Kharin (2013), Kharin (2005), Kharin and Shlyk (2009). Therefore, small-parametric or parsimonious models are developed to overcome this difficulty. These models are special cases of the s-order Markov chain, but the number of parameters required to determine the one-step transition probability matrix is much less than $D(s)$. Let us give some examples of such parsimonious models: the Markov chain of the order s with r partial connections (see Kharin and Petlitskii 2007), Raftery model (see Raftery 1985), variable length Markov chain (see Buhlmann and Wyner 1999). For example, the conditional probability distribution of the current state of the Markov chain of the order s with r partial connections depends not on all s previous states, but only on r

selected states. This paper is devoted to a new parsimonious model called Markov chain of conditional order proposed by authors in Kharin and Maltsew (2012).

2. Mathematical model

At first let us introduce the notation: $\mathbb{N}$ is the set of positive integers, $N \in \mathbb{N}$, $2 \leq N < \infty$, $A = \{0, 1, \dots, N-1\}$ is the finite state space with N elements; $J_n^m = (j_n, \dots, j_m) \in A^{m-n+1}$, $m \geq n$, is the multiindex (subsequence of indices from a sequence $j_1, j_2, \dots$); $\{x_t \in A : t \in \mathbb{N}\}$ is a homogeneous Markov chain of the order s, $(2 \leq s < \infty)$ with $(s+1)$-dimensional matrix of transition probabilities $P = (p_{J_1^{s+1}})$:

$$p_{J_1^{s+1}} = \mathrm{P}\{x_{t+s} = j_{s+1} | x_{t+s-1} = j_s, \dots, x_t = j_1\},\ J_1^{s+1} \in A^{s+1},\ t \in \mathbb{N};$$

$L \in \{1, 2, \dots, s-1\}$, $K = N^L - 1$ are some positive integers; $Q^{(1)}, \dots, Q^{(M)}$ are M $(1 \leq M \leq K+1)$ different square stochastic matrices of the order N:

$$Q^{(m)} = (q_{i,j}^{(m)}),\ 0 \leq q_{i,j}^{(m)} \leq 1,\ \sum_{j \in A} q_{i,j}^{(m)} \equiv 1,\ i, j \in A,\ 1 \leq m \leq M;$$

$< J_n^m >= \sum\limits_{k=n}^{m} N^{k-n} j_k \in \{0, 1, \dots, N^{m-n+1} - 1\}$ is the numeric representation of the multiindex $J_n^m \in A^{m-n+1}$; $\mathrm{I}\{C\}$ is the indicator function of event C.

The Markov chain $\{x_t \in A : t \in \mathbb{N}\}$ is called the Markov chain of conditional order (see Kharin and Maltsew 2012), if its one-step transition probabilities have the following parsimonious form:

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

where $1 \leq m_k \leq M$, $1 \leq b_k \leq s - L$, $0 \leq k \leq K$, $\min\limits_{0 \leq k \leq K} b_k = 1$; it is assumed that all elements of the set $\{1, 2, \dots, M\}$ occur in the sequence $m_0, \dots, m_K$. The sequence of elements J_{s-L+1}^{s} is called the base memory fragment (BMF) of the random sequence, L is the length of BMF; the value $s_k = s - b_k + 1$ is called the conditional order. Thus the conditional probability distribution of the state x_t at time t depends not on all s previous states, but it depends only on $L+1$ selected states (j_{b_k}, J_{s-L+1}^{s}). Note that if $L = s-1$, $s_0 = s_1 = \dots = s_K = s$, we have the fully-connected Markov chain of the order s. If $M = K+1$, then each transition matrix corresponds to only one value of the BMF, otherwise there exists a common matrix which corresponds to several values of BMF.

Therefore the Markov chain of conditional order is determined by the following parameters:

- unconditional order s of the Markov chain;
- the length of BMF L;
- $K+1$ conditional orders $\{s_k : 0 \leq k \leq K\}$;
- $K+1$ parameters $\{m_k : 0 \leq k \leq K\}$ which determine the transition matrices;
- M stochastic matrices of the order N which are described by $MN(N-1)$ independent parameters.

Hence the transition matrix $P = (p_{J_1^{s+1}})$, $J_1^{s+1} \in A^{s+1}$, of the Markov chain of conditional order is determined by

$$d = 2(N^L + 1) + MN(N-1) \tag{2}$$

independent parameters. For example, we need no more than 66 parameters for the Markov chain of conditional order if $s = 10$, $L = 2$, whereas the fully-connected Markov chain of this order requires $D(s) = 1024$ parameters.

3. Statistical estimators for parameters

In this section we present statistical estimators for parameters of the Markov chain of conditional order. Introduce the notation: $X_1^n \in A^n$ is the observed time series of length n, $\pi^0_{J_1^s} = \mathrm{P}\{x_1 = j_1, \dots, x_s = j_s\}, J_1^s \in A^s$, is the initial probability distribution of the Markov chain of conditional order (1);

$$\nu^s_{l,y}(J_1^l) = \sum_{t=1}^{n-s} \mathrm{I}\{x_{t+s-l-y+1} = j_1, X_{t+s-l+2}^{t+s} = J_2^l\},\ l \geq 2,\ 0 \leq y \leq s-l+1,$$

is frequency of the state $J_1^l \in A^l$ with the time gap of length y between the elements j_1 and J_2^l; $\nu_{s+1}(J_1^{s+1}) = \nu^s_{s+1,0}(J_1^{s+1})$ is frequency of $(s+1)$-tuple J_1^{s+1}.

At first, let us give ergodicity conditions for the Markov chain of conditional order.

Theorem 1. *The Markov chain of conditional order is ergodic if and only if there exists a number $m \in \mathbb{N}, s \leq m < \infty$, such that the following inequality holds:*

$$\min_{J_1^s, J_{1+m}^{s+m} \in A^s} \sum_{J_{s+1}^m \in A^{m-s}} \prod_{i=1}^{m} \sum_{k=0}^{K} \mathrm{I}\{< J_{i+s-L}^{i+s-1} >= k\} q^{(m_k)}_{j_{b_k+i-1}, j_{i+s}} > 0. \tag{3}$$

Proof. Consider the first-order vector-valued Markov chain

$$\{X_t = (x_t, x_{t+1}, \dots, x_{t+s-1}) \in A^s : t \in \mathbb{N}\}$$

with the extended state space like in Doob (1953) which is equivalent to the s-order Markov chain $\{x_t \in A : t \in \mathbb{N}\}$. The transition matrix for X_t has the following form:

$$\bar{P} = (\bar{p}_{J_1^{2s}}),\ J_1^{2s} \in A^{2s},\ \bar{p}_{J_1^{2s}} = \mathrm{I}\{J_2^s = J_{s+1}^{2s-1}\} p_{J_1^s j_{2s}}. \tag{4}$$

According to Kemeny and Snell (1963) the Markov chain X_t is ergodic if and only if there exists a number $m \in \mathbb{N}$, such that the following inequality holds:

$$\min_{J_1^s, J_{1+c}^{s+c} \in A^s} \bar{p}^{(c)}_{J_1^s J_{1+c}^{s+c}} > 0,$$

where $\bar{p}^{(c)}_{J_1^s J_{1+c}^{s+c}}$ is the c-step transition probability from J_1^s to J_{1+c}^{s+c} for the Markov chain X_t. Using properties of probability and definition (1) we come to the criterion (3). Theorem is proved.

In the sequel we will consider ergodic Markov chains. It is known, that the probability distribution of an ergodic Markov chain tends to a stationary probability distribution. The next theorem determines conditions under which the stationary distribution is uniform.

Theorem 2. *If the Markov chain of conditional order is ergodic, then its stationary distribution is uniform if and only if the following equations hold ($k = 0, 1, \dots, K$):*

$$\begin{cases} q_{ij}^{(m_k)} = 1/N, \forall i, j \in A, \quad \text{if } s_k \in \{L+1, \dots, s-1\}, \\ \sum\limits_{i \in A} q_{ij}^{(m_k)} = 1, \forall j \in A \text{ (that is } Q^{(m_k)} \text{ is a doubly stochastic matrix)}, \quad \text{if } s_k = s. \end{cases} \tag{5}$$

Proof. As in the proof of Theorem 1 consider the first-order vector Markov chain X_t. It is known from Borovkov (1998b), that the stationary distribution for X_t is uniform if and only if $\bar{P}$ is a doubly stochastic matrix, that is

$$\sum_{J_1^s \in A^s} \bar{p}_{J_1^{2s}} = 1,\ \forall J_{s+1}^{2s} \in A^s. \tag{6}$$

Define $k =< J_{2s-L}^{2s-1} >$ and transform (6) using (4) and (1):

$$\sum_{J_1^s \in A^s} \bar{p}_{J_1^{2s}} = \sum_{J_1^s \in A^s} \mathbf{I}\{J_2^s = J_{s+1}^{2s-1}\} q^{(m_k)}_{j_{b_k}, j_{2s}} = \sum_{j_1 \in A} q^{(m_k)}_{j_{b_k}, j_{2s}} = 1. \tag{7}$$

If $s_k = s$, then $b_k = 1$ and $\sum\limits_{j_1 \in A} q^{(m_k)}_{j_1, j_{2s}} = 1$. Hence $Q^{(m_k)}$ is a doubly stochastic matrix, and we have the second row in (5). If $s_k < s$, then $b_k > 1$, $\sum\limits_{j_1 \in A} q^{(m_k)}_{j_{b_k}, j_{2s}} = N q^{(m_k)}_{j_{b_k}, j_{2s}} = 1$, and we have the first row in (5). Theorem is proved.

We will use the likelihood function to estimate transition probability matrices $\{Q^{(m_k)}\}$ and conditional orders $\{s_k\}$. In order to build it we have to find n-dimensional probability distribution for the observed time series X_1^n generated by the model (1).

Lemma 1. *The n-dimensional probability distribution ($n > s$) for the Markov chain of conditional order (1) has the following form:*

$$\mathrm{P}\{x_1 = j_1, \dots, x_n = j_n\} = \pi^0_{J_1^s} \prod_{t=s}^{n-1} \sum_{k=0}^{K} \mathrm{I}\{< J_{t-L+1}^t >= k\} q^{(m_k)}_{t-s+b_k, j_{t+1}}, \; j_1, \dots, j_n \in A. \tag{8}$$

Proof. Using theorem on compound probabilities and the Markov property we have:

$$\mathrm{P}\{x_1 = j_1, \dots, x_n = j_n\} = \pi^0(J_1^s) \prod_{t=s}^{n-1} p_{J_{t-s+1}^{t+1}}.$$

Hence, taking into account definition (1), we come to (8). Lemma is proved.

Corollary 1. *The loglikelihood function for the Markov chain of conditional order (1) has the following form:*

$$l_n(X_1^n, \{Q^{(i)}\}, L, \{s_k\}, \{m_k\}) = \ln \pi^0_{X_1^s} +$$
$$+ \sum_{J_0^{L+1} \in A^{L+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\} \nu^s_{L+2, s_k-L-1}(J_0^{L+1}) \ln q^{(m_k)}_{j_0, j_{L+1}}.$$

Now we can construct maximum likelihood estimators (MLEs) for the transition probabilities $\{Q^{(m_k)} : k = 0, \dots, K\}$ and the conditional orders $\{s_k : k = 0, \dots, K\}$.

Theorem 3. *If the true values s, L, $\{s_k : k = 0, \dots, K\}$ and $\{m_k : k = 0, \dots, K\}$ are known, then the MLEs for the one-step transition probabilities $\{q^{(m_k)}_{j_0, j_{L+}}, j_0, j_{L+1} \in A : k = 0, \dots, K\}$ are*

$$\hat{q}^{(m_k)}_{j_0, j_{L+1}} = \begin{cases} \dfrac{\sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+2, g(s_k, L)}(J_0^{L+1})}{\sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+1, g(s_k, L)}(J_0^L)}, \text{ if } \sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+1, g(s_k, L)}(J_0^L) > 0, \\ 1/N, \text{ if } \sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+1, g(s_k, L)}(J_0^L) = 0, \end{cases} \tag{9}$$

where $M_i = \{J_1^L \in A^L : m_{< J_1^L >} = i\}$, $i = 1, \dots, M$, $\bigcup\limits_{i=1}^{M} M_i = A^L$, $g(i, j) = i - j - 1$.

Proof. In order to construct the MLEs we need to solve the following problem:

$$l_n(X_1^n, \{Q^{(i)}\}, L, \{s_k\}, \{m_k\}) \to \max_{\{Q^{(m_k)}\}_{1 \le m_k \le M}},$$
$$\sum_{j_{L+1} \in A} q^{(m_k)}_{j_0, j_{L+1}} = 1, \; j_0 \in A, \; 1 \le m_k \le M.$$

This maximization problem splits into N^{L+1} subproblems ($j_0 \in A, J_1^L \in A^L$):

$$\sum_{j_{L+1}\in A}\sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\}\nu_{L+2,g(s_k,L)}(J_0^{L+1}) \ln q_{j_0,j_{L+1}}^{(m_k)} \to \max_{q_{j_0,j_{L+1}}^{(m_k)}},$$

$$\sum_{j_{L+1}\in A} q_{j_0,j_{L+1}}^{(m_k)} = 1.$$

Solve these subproblems with Lagrange multiplier method and come to the estimators (9). Theorem is proved.

In the rest of the paper we will assume that $M = K+1$, i.e. $K+1$ independent matrices correspond to $K+1$ different values of BMF, and $m_k = k+1, k = 0,1,\ldots,K$. In this case estimators (9) have the following form:

$$\hat{q}_{j_0,j_{L+1}}^{(k+1)} = \begin{cases} \sum\limits_{J_1^L\in A^L} \mathrm{I}\{< J_1^L >= k\}\dfrac{\nu^s_{L+2,g(s_k,L)}(J_0^{L+1})}{\nu^s_{L+1,g(s_k,L)}(J_0^L)}, \text{ if } \nu^s_{L+1,g(s_k,L)}(J_0^L) > 0, \\ 1/N, \text{ if } \nu^s_{L+1,g(s_k,L)}(J_0^L) = 0. \end{cases} \tag{10}$$

We will also use the following notation for transition probabilities and their estimators:

$$q(J_0^{L+1}) = \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\} q_{j_0,j_{L+1}}^{(k+1)}, \ \hat{q}(J_0^{L+1}) = \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\}\hat{q}_{j_0,j_{L+1}}^{(k+1)}.$$

According to Kharin and Maltsew (2011) we construct estimators for the conditional orders $\{s_k\}$.

Theorem 4. *If s and L are known, then the MLEs for conditional orders $\{s_k : k = 0,\ldots,K\}$ are*

$$\hat{s}_k = arg \max_{L+1\le y\le s} \sum_{J_1^L\in A^L} \mathrm{I}\{< J_1^L >= k\} \sum_{j_0,j_{L+1}\in A} \nu^s_{L+2,g(y,L)}(J_0^{L+1}) \ln(\hat{q}_{j_0,j_{L+1}}^{(k+1)}). \tag{11}$$

In order to estimate the order s and the BMF length L we use Bayesian information criterion (BIC) (see Csiszar and Shields 1999):

$$(\hat{s},\hat{L}) = arg \min_{2\le s'\le S_+,\ 1\le L'\le L_+} BIC(s',L'), \tag{12}$$

$$BIC(s',L') = -2 \sum_{J_0^{L'+1}\in A^{L'+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^{L'} >= k\}\nu^{s'}_{L'+2,\hat{g}(s_k,L')}(J_0^{L'+1}) \ln \hat{q}_{j_0,j_{L'+1}}^{(k+1)} + \\ + d\ln(n-s'),$$

where $S_+ \ge 2$, $1 \le L_+ \le S_+ - 1$, are maximal admissible values of s and L respectively, d is the number of independent parameters of the model (1) defined by formula (2).

4. Asymptotic properties of statistical estimators

Let us assume that the Markov chain (1) satisfies the stationarity condition. Define the probability distribution of the l-tuple $X_{t+l-1}^t \in A^l$, $l \in \mathbb{N}$:

$$\pi_l(J_1^l) = \mathrm{P}\{x_t = j_1,\ldots,x_{t+l-1} = j_l\},\ J_1^l \in A^l,\ t = 1,2,\ldots.$$

At first, let us present results on consistency of the constructed statistical estimators from the previous section.

Theorem 5. *If Markov chain of conditional order (1) is stationary, then the statistical estimators (9) are consistent estimators as $n \to \infty$:*

$$\hat{q}_{ij}^{(k+1)} \xrightarrow{\mathrm{P}} q_{ij}^{(k+1)},\ i,j \in A,\ k = 0,\dots,K. \tag{13}$$

Proof. It is known from Basawa and Prakasa Rao (1980) that frequencies of the states for the first-order vector Markov chain X_t (considered in the proof of Theorem 1) tend to the stationary probability distribution as $n \to \infty$:

$$\frac{1}{n-s}\sum_{t=1}^{n-s} \mathrm{I}\{X_t = J_1^s, X_{t+1} = J_2^{s+1}\} \xrightarrow{\mathrm{P}} \pi_{s+1}(J_1^{s+1}),\ J_1^{s+1} \in A^{s+1}.$$

Thus we can prove that $\hat{\pi}_{s+1}(J_1^{s+1}) = \nu_{s+1}(J_1^{s+1})/(n-s) \xrightarrow{\mathrm{P}} \pi_{s+1}(J_1^{s+1})$. Then we consider $\nu^s_{L+2,g(s_k,L)}(J_0^{L+1})$ and $\nu^s_{L+1,g(s_k,L)}(J_0^L)$ as sums of the frequencies of $(s+1)$-tuples $\nu_{s+1}(J_1^{s+1})$:

$$\nu^s_{l+1,g(s_k,L)}(J_0^l) = \sum_{I_1^{s+1} \in A^{s+1}(g(s_k,L),J_0^l)} \nu_{s+1}(J_1^{s+1}),\ l \in \{L, L+1\},$$

where $A^{s+1}(y, J_0^l) = \{I_1^{s+1} \in A^{s+1} : i_1 = j_0, I_{y+2}^{y+l} = J_2^l\},\ y = 0,1,\dots.$ So the following convergence holds:

$$\nu^s_{l+2,g(s_k,L)}(J_0^{l+1}) \xrightarrow{\mathrm{P}} \pi_{l+1,g(s_k,L)}(J_0^l) = \mathrm{P}\{x_t = j_0, X_{t+s_k-L}^{t+s_k-L+l-1} = J_1^l\}.$$

Note that $\pi_{L+2,g(s_k,L)}(J_0^{L+1}) = \sum\limits_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\}\pi_{L+1,g(s_k,L)}(J_0^L) q^{(k+1)}_{j_0,j_{L+1}}$; using this equation and theorem on functional transformations of convergent random sequences from Borovkov (1998a), we come to (13). Theorem is proved.

Theorem 6. *Under conditions of Theorem 5 statistical estimators (11) are consistent as $n \to \infty$:*

$$\hat{s}_k \to s_k,\ k = 0,\dots,K+1. \tag{14}$$

Proof. Introduce the notation:

$$I_k(y) = \sum_{j_0,j_{L+1} \in A} \pi_{L+2,g(y,L)}(J_0^{L+1}) \ln \frac{\pi_{L+2,g(y,L)}(J_0^{L+1})}{\pi_{L+1,g(y,L)}(J_0^L)\pi_1(j_{L+1})}, y \in \{L+1,\dots,s\},$$

is the Shannon information on the random symbol x_{L+1} contained in the random symbol x_0 under the fixed BMF $X_1^L = J_1^L$;

$$\bar{I}_k = \sum_{H_1^{s-L} \in A^{s-L}} \sum_{j_{L+1} \in A} \pi_{s+1}(H_1^{s-L} J_1^{L+1}) \ln \frac{\pi_{s+1}(H_1^{s-L} J_1^{L+1})}{\pi_s(H_1^{s-L} J_1^L)\pi_1(j_{L+1})}, y \in \{L+1,\dots,s\},$$

is the Shannon information on the random symbol x_{s+1} contained in the $(s-L)$-tuple X_1^{s-L} under the fixed BMF $X_{s-L+1}^s = J_1^L$;

$$\hat{I}_k(y) = \sum_{j_0,j_{L+1} \in A} \hat{\pi}_{L+2,g(y,L)}(J_0^{L+1}) \ln \frac{\hat{\pi}_{L+2,g(y,L)}(J_0^{L+1})}{\hat{\pi}_{L+1,g(y,L)}(J_0^L)\hat{\pi}_1(j_{L+1})}, y \in \{L+1,\dots,s\},$$

is the plug-in statistical estimator for $I_k(y)$. At first, note that

$$arg \max_{L+1 \le y \le s} \sum_{j_0,j_{L+1} \in A} \nu^s_{L+2,g(y,L)}(J_0^{L+1}) \ln(\hat{q}^{(k+1)}_{j_0,j_{L+1}}) = arg \max_{L+1 \le y \le s} \hat{I}_k(y), \tag{15}$$

where $< J_1^L >= k$. The second statement we need to prove the theorem, is the following:

$$I_k(s_k) = \bar{I}_k. \tag{16}$$

Using (16) and properties of Shannon information we can show that $I_k(s_k) \geq I_k(y)$, $\forall y \neq s_k$. Thus applying the first continuity theorem from Borovkov (1998a) and the equation (15) we come to (14). Theorem is proved.

Theorem 7. *Under conditions of Theorem 5 statistical estimators (12) are consistent as $n \to \infty$:*

$$(\hat{s}, \hat{L}) \xrightarrow{\mathrm{P}} (s, L).$$

Proof. Let $\pi_{l,y}(J_1^l) = \mathrm{P}\{x_t = j_1, X_{t+y+1}^{t+y+l-1} = J_2^l\}$, $l \geq 2$, $y \geq 0$. Then $q_{j_0,j_{L+1}}^{(k+1)} = \dfrac{\pi_{L+2,g(s_k,L)}(J_0^{L+1})}{\pi_{L+1,g(s_k,L)}(J_0^L)}$, where $< J_1^L >= k$. Note that if $X_1^{L'} = J_1^{L'}$ is fixed, then $-\sum\limits_{j_0,j_{L'} \in A} \pi_{L'+2,y}(J_0^{L'+1}) \ln \dfrac{\pi_{L'+2,y}(J_0^{L'+1})}{\pi_{L'+1,y}(J_0^{L'})}$ is a conditional entropy $H_{J_1^{L'},y}\{x_{L'+1}|x_0\}$ of $x_{L'+1}$ given x_0. Using asymptotic properties of the estimators (10) and (11) it is easy to show that for $n \to \infty$ the following asymptotics holds:

$$-\frac{1}{n} \sum_{J_0^{L'+1} \in A^{L'+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^{L'} >= k\} \nu_{L'+2,g(\hat{s}_k,L')}^{s'}(J_0^{L'+1}) \ln \frac{\nu_{L'+2,g(\hat{s}_k,L')}^{s'}(J_0^{L'+1})}{\nu_{L'+1,g(\hat{s}_k,L')}^{s'}(J_0^{L'})} \xrightarrow{\mathrm{P}}$$

$$\xrightarrow{\mathrm{P}} \sum_{J_1^{L'} \in A^{L'}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^{L'} >= k\} H_{J_1^{L'},g(y_k,L')}\{x_{L'+1}|x_0\},$$

where $L' + 1 \leq y_k \leq s'$. Using properties of entropy and methods described in Csiszar and Shields (1999) we can prove that $\mathrm{P}\{(\hat{s}, \hat{L}) \in \{[2, S+] \times [1, L_+]\} \setminus \{(s, L)\}\} \xrightarrow{\mathrm{P}} 0$ at $n \to \infty$. Theorem is proved.

Now let us analyze the asymptotic normality property for estimators (10). Theorem 8 establishes asymptotic probability distribution of the normalized deviations of the statistical estimators for transition probabilities:

$$\bar{q}(J_0^{L+1}) = \sqrt{n-s}(\hat{q}(J_0^{L+1}) - q(J_0^{L+1}))\ , J_0^{L+1} \in A^{L+2}.$$

Theorem 8. *Under conditions of Theorem 5 as $n \to \infty$ the normalized deviations $\{\bar{q}(J_0^{L+1}) : J_0^{L+1} \in A^{L+2}\}$ have joint asymptotically normal probability distribution with zero mean and covariance matrix $\Sigma_q = \Sigma_q(H_0^{L+1}, J_0^{L+1})$, H_0^{L+1}, $J_0^{L+1} \in A^{L+2}$:*

$$\Sigma_q(H_0^{L+1}, J_0^{L+1}) = \mathrm{I}\{H_0^L = J_0^L\} q(H_0^{L+1}) \frac{\mathrm{I}\{h_{L+1} = j_{L+1}\} - q(H_0^L j_{L+1})}{\pi(H_0^L)}. \tag{17}$$

Proof. Let us give only a scheme of the proof. Complete proof can be found in Kharin and Maltsew (2012). The theorem is proved using asymptotic normality property for frequencies $\nu_{s+1}(J_1^{s+1})$ from Kharin and Petlitskii (2007). We represent the estimator $\bar{q}(J_0^{L+1})$ as a function of these frequencies. Therefore using the third continuity theorem from Borovkov (1998a) we can establish asymptotic normality property for estimators (10) and come to (17). Theorem is proved.

5. Statistical testing of hypotheses on the values of $\{Q^{(k)}\}$

Using the results of Section 4 let us construct a statistical test for two hypotheses:

$$H_0 = \{Q^{(1)} = Q_0^{(1)}, \ldots, Q^{(K+1)} = Q_0^{(K+1)}\}, H_1 = \bar{H}_0, \tag{18}$$

where $Q_0^{(1)}, \dots, Q_0^{(K+1)}$ are some fixed $K+1$ stochastic matrices of the order N.

For the decision making we will use the following statistic:

$$\rho = \rho(n) = \sum_{J_0^L \in A^{L+1}} \sum_{j_{L+1} \in Q(J_0^L)} \bar{q}_0^2(J_0^{L+1}) \pi_{L+1}(J_0^L)/q(J_0^{L+1}),$$

$$Q(J_0^L) = \{j_{L+1} \in A : q(J_0^{L+1}) > 0\},$$

where $\bar{q}_0^2(J_0^{L+1}) = \sqrt{n-s}(\hat{q}(J_0^{L+1}) - q_0(J_0^{L+1}))$.

Theorem 9. *Under conditions of Theorem 5 as $n \to \infty$ the probability distribution of the random variable $\rho(n)$ tends to the standard χ^2-distribution with u degrees of freedom,*

$$u = \sum_{J_0^L \in A^{L+1}} \left(|Q^{(}J_0^L)| - 1 \right).$$

Proof. Let us give only a scheme of the proof. Complete proof can be found in Kharin and Maltsew (2012). Since normalized deviations $\{\bar{q}(J_0^{L+1}) : J_0^{L+1} \in A^{L+1}\}$ have the joint asymptotically normal distribution according to Theorem 8, we can establish the probability distribution of $\rho(n)$ using the theorem on quadratic forms for multidimensional Gaussian vectors and the second continuity theorem from Borovkov (1998a). Theorem is proved.

Now we can construct the statistical test for the hypotheses (18) based on the statistic $\rho(n)$:

$$\text{accept the hypothesis} \begin{cases} H_0, \ if \ \rho(n) \le \Delta, \\ H_1, \ if \ \rho(n) > \Delta, \end{cases} \tag{19}$$

where $\Delta = G_u^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the standard χ^2-distribution with u degrees of freedom, $\alpha \in (0,1)$ is the given significance level.

Corollary 2. *Under conditions of Theorem 5 as $n \to \infty$ the asymptotic size of the test (19) is equal to the given significance level $\alpha \in (0,1)$:*

$$\alpha_n = \mathrm{P}\{\rho(n) > \Delta | H_0\} \xrightarrow[n \to \infty]{} \alpha.$$

Let us consider now the alternative hypothesis of the following special type:

$$H_{1n} = \{Q^{(1)} = Q_1^{(1)}, \dots, Q^{(K+1)} = Q_1^{(K+1)}\}, \tag{20}$$

$$Q_1^{(k)} = Q_0^{(k)} + \frac{1}{\sqrt{n-s}} \gamma^{(k)}, \ \gamma^{(k)} = (\gamma_{i,j}^{(k)}), i,j \in A, k = 1, \dots, K+1,$$

where $\{\gamma^{(k)}\}$ are some fixed square matrices of the order N, such that $\sum_{j \in A} \gamma_{i,j}^{(k)} = 0$, $\sum_{i,j \in A} (\gamma_{i,j}^{(k)})^2 > 0$. Formula (20) means that the alternative hypothesis H_{1n} tends to the null hypothesis H_0 as $n \to \infty$; such a family of hypotheses $\{H_{1n} : n = 1, 2, \dots\}$ is called the family of contigual hypotheses (see Roussas 1972). For this case we can obtain the asymptotic power of the test (19). The next theorem is proved by analogy with Theorem 9. Complete proof is given in Kharin and Maltsew (2012).

Theorem 10. *If the Markov chain of conditional order (1) is stationary and the contigual family of alternatives (20) holds, then as $n \to \infty$ the probability distribution of the random variable $\rho(n)$ tends to the noncentral χ^2-distribution with u degrees of freedom and the noncentrality parameter λ:*

$$\lambda = \sum_{\substack{J_0^L \in A^{L+1}, \\ j_{L+1} \in Q(J_0^L)}} \frac{\pi_{L+1}(J_0^L)}{q(J_0^{L+1})} \gamma^2(J_0^{L+1}),$$

where $\gamma(J_0^{L+1}) = \sum\limits_{k=1}^{K+1} \mathrm{I}\{< J_1^L >= k\}\gamma^{(k)}_{j_0, j_{L+1}}$.

Corollary 3. *Under conditions of Theorem 9 the power of the test (19) as $n \to \infty$ tends to the limit:*

$$w = 1 - G_{u,\lambda}(G_u^{-1}(1-\alpha)), \tag{21}$$

where $G_{u,\lambda}$ is the distribution function of the noncentral χ^2-distribution with u degrees of freedom and the noncentrality parameter λ and $\alpha \in \{0, 1\}$ is the given significance level.

Let us note that the power doesn't tend to 1 because the alternative hypothesis H_{1n} tends to the null hypothesis as $n \to \infty$.

6. Computer experiments on hypothesis testing

Simulated data. At first, we evaluate the test (19) performance for contigual alternatives (20) in two series of computer experiments by the following scheme: $U = 1000$ realizations of the Markov chain of conditional order were simulated according to (1). Parameters of the model: $N = 2, A = \{0, 1\}, s = 8, L = 2,\ M = 4, s_0 = 8, s_1 = 6, s_2 = 8, s_3 = 3$. The length of the time series $n \in \{1000, 1500, \ldots, 20000\}$. **In the first series** of experiments the transition probabilities were chosen randomly for the null hypothesis H_0. **In the second series** of experiments transition probabilities were chosen randomly to provide alternative hypothesis H_1. In both series the frequency of the decision "accept the hypothesis H_1" was calculated at the fixed value of n:

$$\nu_\rho = \frac{1}{U}\sum_{u=1}^{U} \mathrm{I}\{\rho_u(n) > \Delta\},$$

where $\rho_u(n)$ is the value of $\rho(n)$ calculated by the u-th realization. In the first series ν_ρ is the estimator of the error I probability, we will denote it $\hat{\alpha}$. In the second series ν_ρ is the estimator of the power, we will denote it $\hat{w}$. Results for the first series of experiments are presented in Figure 1; results for the second series of experiments are presented in Figure 2. On both figures horizontal axis corresponds to the time series length n, vertical axis corresponds to the value of ν_ρ; in both cases $\alpha = 0.05$. Solid line in Figure 1 plots the significance level α. Solid line in Figure 2 plots the theoretical power (21) of the test. As we can see, theoretical values of α and w are close enough to their experimental values $\hat{\alpha}$, $\hat{w}$ respectively which are indicated by dark circles.

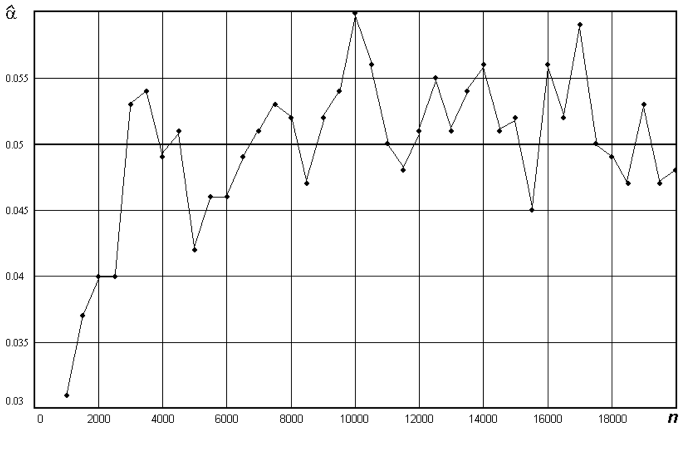

Figure 1: Dependence $\hat{\alpha}$ on n.

Real data. The real data we used is a genetic sequence from the human DNA. Splitting of genes into coding segments (exons) and noncoding segments (introns) is an important problem

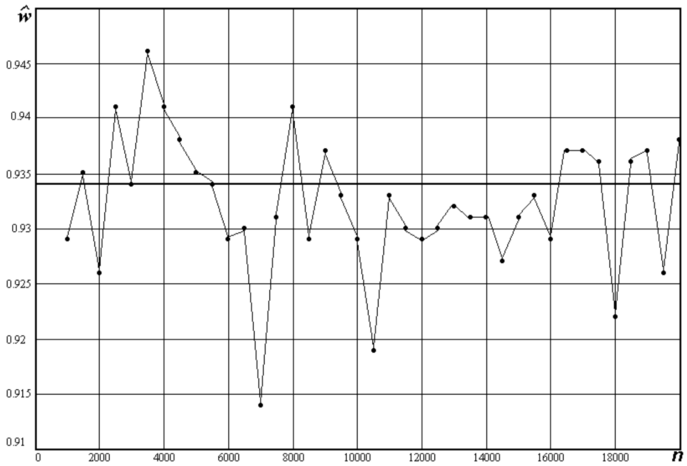

Figure 2: Dependence $\hat{w}$ on n.

in bioinformatics, and fitting a stochastic model for genetic sequence is a fruitful approach to this problem decsribed in Burge and Karlin (1997).

The sequence of introns from the human gene HSHMG17G taken from "Bioinformatics and genomics" (http://genome.crg.es/) was analyzed. The length of the sequence $n = 6922$, $S_+ \leq 6$, the size of the state space A is 4 (0 corresponds to nucleotide A, 1 to C, 2 to G, 3 to T). We used in computer experiments the following three Markov chain models: fully-connected s-order Markov chain (MC(s)), the Markov chain of order s with r partial connections (MC(s, r)) and the Markov chain of conditional order with BMF length L (MCCO(s, L)). For each model the value of BIC was calculated. Results are presented in Table 1. Minimum value of BIC is marked by bold type.

Table 1: Values of BIC.

model	BIC	model	BIC	model	BIC
MC(1)	17792.7	MC(4, 3)	18162.9	MCCO(3, 1)	17557.5
MC(2)	17595.7	MC(5, 1)	18108.2	MCCO(4, 1)	17472.6
MC(3)	18293.1	MC(5, 2)	17553.8	MCCO(4, 2)	18205.2
MC(4)	22252.5	MC(5, 3)	18219.8	MCCO(5, 1)	17482.5
MC(5)	39894.1	MC(5, 4)	21896.6	MCCO(5, 2)	18170.6
MC(6)	116798.2	MC(6, 1)	18119.8	MCCO(5, 3)	22616.9
MC(2, 1)	18112.9	MC(6, 2)	17568.9	**MCCO(6, 1)**	**17448.8**
MC(3, 1)	18116.7	MC(6, 3)	18150.0	MCCO(6, 2)	18139.9
MC(3, 2)	17535.8	MC(6, 4)	21849.5	MCCO(6, 3)	22520.2
MC(4, 1)	18123.6	MC(6, 5)	26457.0	MCCO(6, 4)	41618.7
MC(4, 2)	17532.9				

As we can see from Table 1, the most adequate model is the Markov chain of conditional order with parameters: $s = 6$, $L = 1$. Estimators for conditional orders are: $\hat{s}_0 = 4$, $\hat{s}_1 = 3$, $\hat{s}_2 = 3$, $\hat{s}_3 = 6$. Estimates for transition matrices for this MCCO(6, 1) model are:

$$\hat{Q}^{(1)} = \begin{pmatrix} 0.484 & 0.376 & 0.083 & 0.057 \\ 0.463 & 0.405 & 0.085 & 0.047 \\ 0.251 & 0.181 & 0.373 & 0.195 \\ 0.312 & 0.201 & 0.294 & 0.193 \end{pmatrix}, \; \hat{Q}^{(2)} = \begin{pmatrix} 0.372 & 0.485 & 0.040 & 0.103 \\ 0.309 & 0.509 & 0.081 & 0.101 \\ 0.220 & 0.265 & 0.240 & 0.275 \\ 0.216 & 0.329 & 0.108 & 0.347 \end{pmatrix},$$

$$\hat{Q}^{(3)} = \begin{pmatrix} 0.254 & 0.210 & 0.270 & 0.266 \\ 0.170 & 0.370 & 0.285 & 0.175 \\ 0.205 & 0.320 & 0.320 & 0.155 \\ 0.196 & 0.253 & 0.306 & 0.245 \end{pmatrix}, \ \hat{Q}^{(4)} = \begin{pmatrix} 0.201 & 0.181 & 0.331 & 0.287 \\ 0.099 & 0.326 & 0.276 & 0.299 \\ 0.125 & 0.230 & 0.342 & 0.303 \\ 0.125 & 0.230 & 0.342 & 0.303 \\ 0.193 & 0.206 & 0.215 & 0.386 \end{pmatrix}.$$

Let us note that the values of BIC close to the minimum are obtained for MCCO(4, 1) and MCCO(5, 1). These two models describe similar dependence to MCCO(6, 1), but they have shorter memory depth. Thus MCCO(6, 1) is chosen as the most adequate model, because the number of parameters for all three models is the same.

7. Conclusion

In this paper we consider a new parsimonious model for discrete-valued time series called Markov chain of conditional order. Probabilistic and statistical properties of the model are established. Ergodicity conditions and conditions under which the stationary probability distribution is uniform are found. Statistical estimators for parameters are constructedwhich and their consistency is proved. Asymptotic probability distribution of the estimators for the transition one-step probabilities is found. Statistical test for the values of transition matrices is constructed and its asymptotic power for contigual alternatives is evaluated. Computer experiments on simulated time series and on real DNA sequences are conducted.

References

Basawa I, Prakasa Rao B (1980). *Statistical Inference for Stochastic Processes.* Academic Press, London.

Borovkov A (1998a). *Mathematical Statistics.* Gordon and Breach, New York.

Borovkov A (1998b). *Probability Theory.* Gordon and Breach, Amsterdam.

Buhlmann P, Wyner A (1999). “Variable Length Markov Chains.” *The Annals of Statistics*, **27**(2), 480–513.

Burge C, Karlin S (1997). “Prediction of Complete Gene Structures in Human Genomic DNA.” *J. Mol. Biol.*, **268**(1), 78–94.

Ching W (2004). “High-order Markov Chain Models for Categorical Data Sequences.” *Naval Research Logistics*, **51**, 557–574.

Csiszar I, Shields P (1999). “Consistency of the BIC Order Estimator.” *Electronic research announcements of the American mathematical society*, **5**, 123–127.

Doob J (1953). *Stochastic Processes.* Wiley, New York.

Kemeny J, Snell J (1963). *Finite Markov Chains.* D. Van Nostrand Company, Princeton NJ.

Kharin A (2005). “Robust Bayesian Prediction Under Distrtions of Prior and Conditional Distributions.” *Journal of Mathematical Sciences*, **126**(1), 992–997.

Kharin A (2013). “Robustness of Sequential Testing of Hypotheses on Parameters of M-valued Random Sequences.” *Journal of Mathematical Sciences*, **189**(6), 924–931.

Kharin A, Shlyk P (2009). “Robust Multivariate Bayesian Forecasting Under Functional Distortions in the Chi-square metric.” *Journal of Statistical Planning and Inference*, **139**, 3842–3846.

Kharin Y, Maltsew M (2011). “Algorithms for Statistical Analysis of Markov Chain with Conditional Memory Depth.” *Informatics*, **1**, 34–43(in Russian).

Kharin Y, Maltsew M (2012). “Hypothesis Testing for Parameters of the Markov Chain of Conditional Order.” *Proceedings of the National Academy of Sciences of Belarus. Series of physical-mathematical sciences*, **3**, 5–12 (in Russian).

Kharin Y, Petlitskii A (2007). “A Markov Chain of Order s with r Partial Connections and Statistical Inference on its Parameters.” *Discrete Mathematics and Applications*, **17**(3), 295–317.

Li Y, Dong Y, Zhang H, Zhao H, Shi H, Zhao X (2010). “Spectrum Usage Prediction Based on High-order Markov Model for Cognitive Radio Networks.” *10th IEEE International Conference on Computer and Information Technology*, pp. 2784–2788.

Raftery A (1985). “A Model for High-order Markov Chains.” *J. Royal Statistical Society*, **B 47**, 528–539.

Roussas G (1972). *Contiguity of Probability Measures: Some Applications in Statistics*. University Press, Cambridge.

Waterman M (1999). *Mathematical Methods for DNA Sequences*. Chapman and Hall/CRC, Boca Raton, Florida.

Affiliation:

Yuriy Kharin
Department of Mathematical Modeling and Data Analysis
Belarusian State University
Independence av. 4
220030 Minsk, Belarus
E-mail: Kharin@bsu.by

Mikhail Maltsau
Research Institute for Applied Problems of Mathematics and Informatics
Belarusian State University
Independence av. 4
220030 Minsk, Belarus
E-mail: Maltsew@bsu.by

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 43/3-4 *Submitted:* 2013-11-25
June 2014 *Accepted:* 2014-03-10