ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 21, Number 1, June 2017
Available online at http://acutm.math.ut.ee

Statistical analysis of high-order Markov dependencies

Yu. S. Kharin and M. V. Maltsew

Abstract. The paper deals with parsimonious models of integer valued time series. Such models are special cases of high-order Markov chain with a small number of parameters. Two new parsimonious models are presented. The first is Markov chain of order s with r partial connections, and the second model is called Markov chain of conditional order. Theoretical results on probabilistic properties and statistical inferences for these models are given.

1. Introduction

An universal model for real-world processes with discrete time t, finite state space $A = \{0, 1, \dots, N-1\}$, $2 \le N < +\infty$, and stochastic dependence of high order $s \gg 1$ (in genetics, computer networks, financial markets, meteorology, and other fields) is the order s homogeneous Markov chain ($MC(s)$) x_t on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ determined by an $(s+1)$-dimensional matrix of one-step transition probabilities

$$P = (p_{j_1,\dots,j_{s+1}}),\ p_{j_1,\dots,j_{s+1}} = \mathbf{P}\{x_{t+1} = j_{s+1} \,|\, x_t = j_s, \dots, x_{t-s+1} = j_1\},$$

where $t \ge s$, $j_1, \dots, j_{s+1} \in A$. Unfortunately, the number of independent parameters for the $MC(s)$ increases exponentially with respect to the order s:

$$d_{MC(s)} = N^s(N-1),$$

thus we need data and computational resources of huge size to identify this model.

To avoid this "curse of dimensionality" we propose to use parsimonious (or "small-parametric" [8]) models for $MC(s)$ that are determined by small number of parameters $d \ll d_{MC(s)}$. Three known examples of parsimonious

Received October 3, 2016.
2010 *Mathematics Subject Classification.* 60J10.
Key words and phrases. High-order Markov chain; parsimonious model; estimator; statistical test.
http://dx.doi.org/10.12697/ACUTM.2017.21.06

models are: the Jacobs–Lewis model [6] with $d_{JL} = N+s-1$ parameters; the MTD-model proposed by A. Raftery [13] with $d_{MTD} = N^2 + s - 1$ parameters; the variable length Markov chain model proposed by P. Buhlmann [2]. In this paper we present two new parsimonious models: Markov chain of order s with r partial connections and Markov chain of conditional order.

2. Markov chain with r partial connections

2.1. Definitions. Markov chain $MC(s,r)$ of order s with r partial connections is determined by the following small-parametric form of the matrix P (see [11]):

$$p_{j_1,\dots,j_{s+1}} = q_{j_{m_1},\dots,j_{m_r},j_{s+1}},\ j_1,\dots,j_{s+1} \in A, \tag{1}$$

where $r \in \{1,\dots,s\}$ is the number of connections; $M_r = (m_1,\dots,m_r) \in \mathrm{M}$ is the integer-valued vector with r ordered components $1 = m_1 < m_2 < \dots < m_r \le s$, called the template of connections; M is the set of all admissible patterns M_r; $Q = (q_{j_1,\dots,j_r,j_{r+1}})$ is a stochastic $(r+1)$-dimensional matrix:

$$0 \le q_{j_1,\dots,j_r,j_{r+1}} \le 1,\ \sum_{j_{r+1}\in A} q_{j_1,\dots,j_r,j_{r+1}} = 1,\ j_1,\dots,j_r \in A.$$

Formula (1) means that the conditional probability distribution of the future state x_{t+1} depends not on all s previous states, but it depends only on r selected states. We need $d = N^r(N-1)$ parameters to completely determine $MC(s,r)$.

Let us now present probabilistic properties and construct statistical estimators for parameters of the model.

2.2. Probabilistic properties of MC(s,r). We denote by $J_n^m = (j_n,\dots,j_m) \in A^{m-n+1}$, $m \ge n$, the multiindex (subsequence of indices from a sequence $j_1, j_2, \dots \in A$).

Theorem 1. *The $MC(s,r)$ defined by (1) is an ergodic Markov chain if and only if there exists $i \in \mathbb{N}$ such that*

$$\min_{J_1^s, J_{s+i+1}^{2s+i} \in A^s} \sum\nolimits_{J_{s+1}^{s+i} \in A^i} \prod_{k=1}^{s+i} q_{j_{k+m_1-1},\dots,j_{k+m_r-1},j_{k+s}} > 0.$$

Stationary probability distribution $(\pi^*_{J_1^s})_{J_1^s \in A^s}$ satisfies the equations

$$\pi^*_{J_2^{s+1}} = \sum\nolimits_{j_1 \in A} \pi^*_{J_1^s}\, q_{j_{m_1},\dots,j_{m_r},j_{s+1}},\ J_1^{s+1} \in A^{s+1}.$$

Proof. Construct 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 [5], 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 form

$$\bar{P} = (\bar{p}_{J_1^{2s}}),\ J_1^{2s} \in A^{2s},\ \bar{p}_{J_1^{2s}} = \mathrm{I}\{J_2^s = J_{s+1}^{2s-1}\} p_{j_1,\ldots,j_s,j_{2s}}.$$

According to [7] the Markov chain X_t is ergodic if and only if there exists a number $c \in \mathbb{N}$, 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 making some transformations we get the required result. □

Corollary 1. *Assume that the MC(s,r) is a stationary Markov chain. The stationary probability distribution has the multiplicative form*

$$\pi^*_{J_1^s} = \prod_{i=1}^{s} \pi^*_{j_i},\ J_1^s \in A^s,$$

if and only if, for any $J_2^{r+1} \in A^r$, we have

$$\pi^*_{j_{r+1}} = \sum\nolimits_{j_1 \in A} \pi^*_{j_1}\, q_{J_1^{r+1}},$$

and the normalizing condition $\sum_{j \in A} \pi^*_j = 1$ holds.

2.3. Statistical inferences on Q. Introduce the notation:

$$X_1^n = (x_1, \ldots, \quad x_n) \in A^n$$

is a realization of the MC(s,r) of the length n that is used for construction of statistical inferences;

$$F\big(J_i^{i+s-1}; M_r\big) = (j_{i+m_1-1}, \ldots, j_{i+m_r-1})$$

is the selector-function of the r-th order;

$$\delta_{J_1^k, I_1^k} = \prod_{l=1}^{k} \delta_{j_l, i_l}$$

is the Kronecker symbol for $J_1^k,\ I_1^k \in A^k$;

$$\nu_{J_1^{r+1}}(X_1^n; M_r) = \sum\nolimits_{t=1}^{n-s} \delta_{F\left(X_t^{t+s-1}; M_r\right), J_1^r} \delta_{x_{t+s}, j_{r+1}} \tag{2}$$

is the frequency statistic of the MC(s,r) for a connection template $M_r \in \mathrm{M}$;

$$\mu_{J_1^{r+1}}(M_r) = \mathbf{P}\big\{F\big(X_t^{t+s-1}; M_r\big) = J_1^r, x_{t+s} = j_{r+1}\big\},\ 1 \le t \le n-s, \tag{3}$$

is the probability distribution of the $(r+1)$-tuple; the dot used instead of any index means summation on all its values: $\mu_{J_1^r \cdot}(M_r) = \sum\limits_{j_{r+1} \in A} \mu_{J_1^{r+1}}(M_r)$.

Theorem 2. *Let the connection template M_r be known. The maximum likelihood estimator (MLE) for the matrix Q is*

$$\hat{Q} = (\hat{q}_{J_1^{r+1}})_{J_1^{r+1}\in A^{r+1}},$$

$$\hat{q}_{J_1^{r+1}} = \begin{cases} \hat{\mu}_{J_1^{r+1}}(M_r)/\hat{\mu}_{J_1^r.}(M_r) & \text{if } \hat{\mu}_{J_1^r.}(M_r) > 0, \\ 1/N & \text{if } \hat{\mu}_{J_1^r.}(M_r) = 0, \end{cases} \tag{4}$$

where

$$\hat{\mu}_{J_1^{r+1}}(M_r) = \nu_{J_1^{r+1}}(X_1^n; M_r)/(n-s)$$

is the frequency estimator for the probability $\mu_{J_1^{r+1}}(M_r)$, $J_1^{r+1} \in A^{r+1}$, $M_r \in$ M.

Proof. Estimators (4) are the solution of the following maximization problem:

$$l(\hat{Q}, M_r) \to \max_{\hat{Q}}, \quad \sum_{j_{r+1}\in A} \hat{q}_{j_1,\dots,j_r,j_{r+1}} = 1, \ j_1, \dots, j_r \in A,$$

where $l(\hat{Q}, M_r)$ is the loglikelihood function for the Markov chain with partial connections:

$$l(Q, M_r) = \ln \pi_{X_1^s} + \sum_{J_1^{r+1}\in A^{r+1}} \nu_{J_1^{r+1}}(X_1^n; M_r) \ln(q_{j_1,\dots,j_{r+1}}), \tag{5}$$

$\pi_{X_1^s}$ is the initial probability distribution. □

Theorem 3 (see [11])**.** *For the stationary MC(s, r) the statistics $\{\hat{q}_{J_1^{r+1}} : J_1^{r+1} \in A^{r+1}\}$, defined by (4), are asymptotically ($n \to \infty$) unbiased and consistent estimators with covariances*

$$\mathrm{Cov}\{\hat{q}_{J_1^{r+1}}, \hat{q}_{K_1^{r+1}}\} = \sigma^{\hat{q}}_{J_1^{r+1},K_1^{r+1}}/(n-s) + \mathcal{O}(1/n^2),$$

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

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

The consistent statistical test for the hypotheses H_0: $Q = Q^0$ of the matrix $Q^0 = (q^0_{J_1^{r+1}})_{J_1^{r+1}\in A^{r+1}}$, against $H_1 = \neg H_0$ consists of the following steps.

1. Computation of the statistics $\nu_{J_1^{r+1}}(X_1^n; M_r)$, $J_1^{r+1} \in A^{r+1}$, by (2).

2. Computation of the statistic

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

where $D_{J_1^r} = \{j_{r+1} \in A : q^0_{J_1^{r+1}} > 0\}$.

3. Computation of the P-value: $P = 1 - G_U(\rho)$, where $G_U(\cdot)$ is the probability distribution function of the standard χ^2 distribution with $U = \sum_{J_1^r \in A^r} (|D_{J_1^r}| - 1)$ degrees of freedom.

4. The decision rule (ε — asymptotic significance level): if $P \geq \varepsilon$, then to stay with the hypothesis H_0 is true; otherwise, the alternative H_1 is true.

2.4. Statistical estimation of the connection template M_r. Introduce the notation:

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

is the conditional entropy of the future symbol $x_{t+s} \in A$ relative to the past derived by the selector

$$F\left(X_t^{t+s-1};\, M_r\right) \in A^r,\ M_r \in \mathrm{M};$$

$\hat{H}(M_r)$ is the "plug-in" estimator of the conditional entropy generated by substitution in (3) estimators $\hat{\mu}_{J_1^{r+1}}(M_r)$, $J_1^{r+1} \in A^{r+1}$, instead of true probabilities $\mu_{J_1^{r+1}}(M_r)$.

Theorem 4. *If the order s and the number of connections r are known, then the maximum likelihood estimator for the true connection template M_r is expressed in terms of the conditional entropy*

$$\hat{M}_r = \arg\min_{M_r \in \mathrm{M}} \hat{H}(M_r). \tag{6}$$

Proof. We get estimator (6) by maximization of the loglikelihood function (5) introduced in Theorem 2. □

Theorem 5 (see [11])**.** *If $MC(s, r)$ is stationary, then the estimator $\hat{M}_r$ defined by* (6)*, at $n \to \infty$, is consistent:*

$$\hat{M}_r \xrightarrow{\mathbf{P}} M_r.$$

2.5. Statistical estimation of the order s and the number of connections r. Let $s \in [s_-, s_+]$, $r \in [r_-, r_+]$, $1 \leq s_- < s_+ < \infty$, $1 \leq r_- < r_+ < s_+$. Estimating the order and the number of connections by maximum likelihood method leads to a problem known as overfitting [3]. Therefore for estimation s and r we use the Bayesian Information Criterion [4], which in our case has the form:

$$BIC(s, r) = 2(n - s)\hat{H}(\hat{M}_r) + U \ln(n - s)\,,$$

where

$$U = \sum_{J_1^r \in A^r} (|G_{J_1^r}| - 1 + \delta_{\hat{\mu}_{J_1^r.}(\hat{M}_r),0}),\ G_{J_1^r} = \{j_{r+1} \in A : \hat{\mu}_{J_1^{r+1}}(\hat{M}_r) > 0\}.$$

Statistical estimators for s and r are determined by minimization:

$$(\hat{s}, \hat{r}) = arg \min_{s_- \le s' \le s_+,\ r_- \le r' \le r_+} BIC(s', r'). \tag{7}$$

Theorem 6 (see [11]). *If $MC(s, r)$ is stationary, then the BIC-estimators $\hat{r}$, $\hat{s}$ defined by (7), at $n \to \infty$, are consistent.*

3. Markov chain of conditional order

3.1. Definitions. Let us introduce the notation: $L \in \{1, 2, \dots, s-1\}$ is some positive integer, $K = N^L - 1$; $Q^{(1)}, \dots, Q^{(M)}$ are M $(1 \le M \le K+1)$ different square stochastic matrices of the order N:

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

$< J_n^m >= \sum_{k=n}^{m} N^{k-n} j_k \in \{0, 1, \dots, N^{m-n+1} - 1\}$ is the numeric representation of the multiindex, $J_n^m \in A^{m-n+1}$; $\mathrm{I}\{C\}$ is the indicator function of the event C. Further $1 \le m_k \le M$, $1 \le b_k \le s - L$, $0 \le k \le K$. It is assumed that sequences $\{m_k\}$ and $\{b_k\}$ are fixed, $\min_{0 \le k \le K} b_k = 1$ and all elements of the set $\{1, 2, \dots, M\}$ occur in the sequence $m_0, \dots, m_K$.

The Markov chain $\{x_t \in A : t \in \mathbb{N}\}$ is called the Markov chain of conditional order (MCCO(s, L)) (see [10]), if its one-step transition probabilities have the following parsimonious form:

$$p_{J_1^{s+1}} = \sum_{k=0}^{K} \mathrm{I}\{< J_{s-L+1}^s >= k\} q_{j_{b_k}, j_{s+1}}^{(m_k)}. \tag{8}$$

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 of Markov chain. Thus the conditional probability distribution of the state x_{t+1} at time point $t+1$ 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 $L = s - 1$, $s_0 = s_1 = \cdots = s_K = s$, we have the fully-connected Markov chain of the order s: $MC(s)$. 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.

Hence the transition matrix P of the Markov chain of conditional order is determined by

$$d = 2(N^L + 1) + MN(N - 1) \tag{9}$$

independent parameters. For example, for $N = 2$ 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 $d_{MC(s)} = 1024$ parameters.

3.2. Probabilistic properties of MCCO. The following theorem, which is proved similarly to Theorem 1, gives ergodicity conditions for the Markov chain of conditional order.

Theorem 7. *The Markov chain of conditional order is ergodic if and only if there exists a number $m \in \mathbb{N}$, $s \le m < \infty$, such that the following inequality holds:*

$$\min_{J_1^s, J_{1+m}^{s+m} \in A^s} \sum_{J_{s+1}^m \in A^{m-s}} \prod_{i=1}^{m} \sum_{k=0}^{K} \mathrm{I}\{< J_{i+s-L}^{i+s-1} >= k\} q^{(m_k)}_{j_{b_k+i-1}, j_{i+s}} > 0. \tag{10}$$

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

$$\begin{cases} q_{ij}^{(m_k)} = 1/N, \forall i, j \in A & \text{if } s_k \in \{L+1, \ldots, s-1\}, \\ \sum\limits_{i \in A} q_{ij}^{(m_k)} = 1, \forall j \in A & \text{if } s_k = s. \end{cases} \tag{11}$$

Proof. As in the proof of Theorem 1, consider the first-order vector Markov chain X_t. The stationary distribution for X_t is uniform if and only if transition matrix $\bar{P}$ is a doubly stochastic matrix, that is

$$\sum_{J_1^s \in A^s} \bar{p}_{J_1^{2s}} = 1, \ \forall J_{s+1}^{2s} \in A^s. \tag{12}$$

Define $k =< J_{2s-L}^{2s-1} >$ and transform (12):

$$\sum_{J_1^s \in A^s} \bar{p}_{J_1^{2s}} = \sum_{J_1^s \in A^s} \mathbf{I}\{J_2^s = J_{s+1}^{2s-1}\} q^{(m_k)}_{j_{b_k}, j_{2s}} = \sum_{j_1 \in A} q^{(m_k)}_{j_{b_k}, j_{2s}} = 1. \tag{13}$$

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 (11). If $s_k < s$, then $1 < b_k \le s - L$ and $q^{(m_k)}_{j_{b_k}, j_{2s}}$ in sum (13) does not depend on j_1: $1 = \sum\limits_{j_1 \in A} q^{(m_k)}_{j_{b_k}, j_{2s}} = N q^{(m_k)}_{j_{b_k}, j_{2s}}$, and we have the first row in (11). □

3.3. Statistical inferences on transition probabilities. Let us now construct 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, $\pi^0_{J_1^s} = \mathrm{P}\{x_1 = j_1, \ldots, x_s = j_s\}$, $J_1^s \in A^s$, is the initial probability distribution of the Markov chain of conditional order (8);

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

is the frequency of the state $J_1^l \in A^l$ with the time gap of length y between the values j_1 and J_2^l;

$$\nu_{s+1}(J_1^{s+1}) = \nu^s_{s+1,0}(J_1^{s+1})$$

is the frequency of $(s+1)$-tuple J_1^{s+1}.

Let us construct now maximum likelihood estimators (MLEs) for the matrices of transition probabilities $\{Q^{(m_k)} : k = 0, \ldots, K\}$.

The loglikelihood function for the Markov chain of conditional order has the form

$$l_n(X_1^n, \{Q^{(i)}\}, L, \{s_k\}, \{m_k\}) = \ln \pi_{X_1^s} + \\ + \sum_{J_0^{L+1} \in A^{L+2}} \sum_{k=0}^{K} \mathrm{I}\{< J_1^L >= k\} \nu^s_{L+2, s_k - L - 1}(J_0^{L+1}) \ln q^{(m_k)}_{j_0, j_{L+1}}. \tag{14}$$

Theorem 9. *If the true values s, L, $\{s_k : k = 0, \ldots, K\}$ and $\{m_k : k = 0, \ldots, K\}$ are known, then the MLEs for the one-step transition probabilities $\{q^{(m_k)}_{j_0, j_{L+1}}, j_0, j_{L+1} \in A : k = 0, \ldots, K\}$ are*

$$\hat{q}^{(m_k)}_{j_0, j_{L+1}} = \begin{cases} \dfrac{\sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+2, g(s_k, L)}(J_0^{L+1})}{\sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+1, g(s_k, L)}(J_0^L)} & \text{if } \sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+1, g(s_k, L)}(J_0^L) > 0, \\ 1/N & \text{if } \sum\limits_{J_1^L \in M_{m_k}} \nu^s_{L+1, g(s_k, L)}(J_0^L) = 0, \end{cases} \tag{15}$$

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

Proof. In order to construct the MLEs we need to maximize the loglikelihood funciton $l_n(X_1^n, \{\hat{Q}^{(i)}\}, L, \{s_k\}, \{m_k\})$ with respect to $\hat{Q}^{(m_k)}$, $1 \leq m_k \leq M$, subject to the following equality constraints:

$$\sum_{j_{L+1} \in A} \hat{q}^{(m_k)}_{j_0, j_{L+1}} = 1,\ j_0 \in A,\ 1 \leq m_k \leq M.$$

This maximization problem splits into NM subproblems ($j_0 \in A, J_1^L \in A^L$):

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

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

After solving these subproblems with Lagrange multiplier method we come to the estimators (15). □

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 (15) have the form

$$\hat{q}^{(k+1)}_{j_0,j_{L+1}} = \begin{cases} \sum\limits_{J_1^L\in A^L} \mathrm{I}\{< J_1^L >= k\}\dfrac{\nu^s_{L+2,g(s_k,L)}(J_0^{L+1})}{\nu^s_{L+1,g(s_k,L)}(J_0^L)} & \text{if } \nu^s_{L+1,g(s_k,L)}(J_0^L) > 0, \\ 1/N & \text{if } \nu^s_{L+1,g(s_k,L)}(J_0^L) = 0. \end{cases} \tag{16}$$

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

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

3.4. Statistical estimators for s_k, s, L. Now let us construct estimators for the conditional orders $\{s_k\}$.

Theorem 10. *If s and L are known, then the MLEs for conditional orders $\{s_k : k = 0, \ldots, K\}$ are*

$$\hat{s}_k = arg \max_{L+1\le y\le s} \sum_{J_1^L\in A^L} \mathrm{I}\{< J_1^L >= k\} \sum_{j_0,j_{L+1}\in A} \nu^s_{L+2,g(y,L)}(J_0^{L+1}) \ln(\hat{q}^{(k+1)}_{j_0,j_{L+1}}). \tag{17}$$

Proof. We get estimators (17) by maximization of the loglikelihood function (14). □

In order to estimate the order s and the BMF length L we use Bayesian information criterion (BIC) as in Subsection 2.5 of this paper:

$$(\hat{s}, \hat{L}) = arg \min_{2\le s'\le S_+,\ 1\le L'\le L_+} BIC(s', L'), \tag{18}$$

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

where $l^k_{j_0,j_{L'+1}} = \nu^{s'}_{L'+2,g(\hat{s}_k,L')}(J_0^{L'+1}) \ln \hat{q}^{(k+1)}_{j_0,j_{L'+1}}$, $S_+ \geq 2$, $1 \leq L_+ \leq S_+ - 1$, are maximal admissible values of s and L respectively, d is the number of independent parameters of the model (8) defined by formula (9).

3.5. Asymptotic properties of statistical estimators. Let us assume that the Markov chain (8) satisfies the stationarity condition. Define the probability distribution of the l-tuple $X^t_{t+l-1} \in A^l$, $l \in \mathbb{N}$:

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

It can be proved [10] that at $n \to \infty$ all constructed estimators are consistent:

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

$$\hat{s}_k \to s_k,$$

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

Now let us analyze the asymptotic normality property for estimators (16). Next theorem 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 11 (see [10])**.** *If Markov chain of conditional order (8) is stationary, then 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 $\Sigma_q = \Sigma_q(H_0^{L+1}, J_0^{L+1})$, $H_0^{L+1},\ J_0^{L+1} \in A^{L+2}$:*

$$\Sigma_q(H_0^{L+1}, J_0^{L+1}) = \mathrm{I}\{H_0^L = J_0^L\} q(H_0^{L+1}) \frac{\mathrm{I}\{h_{L+1} = j_{L+1}\} - q(H_0^L j_{L+1})}{\pi(H_0^L)}. \tag{19}$$

Using this result let us construct a statistical test for two hypotheses:

$$H_0 = \{Q^{(1)} = Q_0^{(1)}, \ldots, Q^{(K+1)} = Q_0^{(K+1)}\}, H_1 = \neg H_0, \tag{20}$$

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

For the decision making we will use the statistic

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

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

Theorem 12. *Under conditions of Theorem 11 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,*

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

Proof. Let us give only a scheme of the proof. Complete proof can be found in [9]. 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 11, 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 [1]. □

Now we can construct the statistical test for the hypotheses (20) based on the statistic $\rho(n)$:

$$\text{accept the hypothesis} \begin{cases} H_0 & \text{if} \quad \rho(n) \le \Delta, \\ H_1 & \text{if} \quad \rho(n) > \Delta, \end{cases} \tag{21}$$

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 11 the asymptotic size of the test* (21) *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:

$$H_{1n} = \{Q^{(1)} = Q_1^{(1)}, \dots, Q^{(K+1)} = Q_1^{(K+1)}\}, \tag{22}$$

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

where $\{\gamma^{(k)}\}$ are some fixed square matrices of the order N, such that $\sum\limits_{j\in A} \gamma_{i,j}^{(k)} = 0$, $\sum\limits_{i,j\in A} (\gamma_{i,j}^{(k)})^2 > 0$. Formula (22) 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 [14]. For this case we can obtain the asymptotic power of the test (21). The next theorem is proved similarly to Theorem 12.

Theorem 13. *If the Markov chain of conditional order* (8) *is stationary and the contigual family of alternatives* (22) *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 $\gamma(J_0^{L+1}) = \sum_{k=1}^{K+1} \mathrm{I}\{< J_1^L >= k\} \gamma_{j_0, j_{L+1}}^{(k)}$.

Corollary 3. *Under conditions of Theorem 9 the power of the test* (21) *as $n \to \infty$ tends to the limit*

$$w = 1 - G_{u,\lambda}(G_u^{-1}(1-\alpha)), \tag{23}$$

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 does not tend to 1 because the alternative hypothesis H_{1n} tends to the null hypothesis as $n \to \infty$.

4. Conclusions

Using of high-order Markov chains for modeling of long memory integer valued processes leads to the hard "dimensionality problem", and construction of small-parametric models is necessary for practice. Convenient models for modeling in the indicated situation are the models considered in this paper: the Markov chain with partial connections $\mathrm{MC}(s, r)$ and the Markov chain of conditional order $\mathrm{MCCO}(s, L)$. Probabilistic properties of $\mathrm{MC}(s, r)$, $\mathrm{MCCO}(s, L)$ are investigated, statistical inferences on the model parameters are constructed. Practical implementation of $\mathrm{MCCO}(s, L)$ can be found in [12].

Acknowledgement. The authors thank the anonymous referee for careful reading and valuable suggestions which helped to improve the quality of our paper.

References

[1] A. Borovkov, *Mathematical Statistics*, Gordon and Breach Science Publishers, Amsterdam, 1998.

[2] P. Buhlmann, *Model selection for variable length Markov chains and tuning the context algorithm*, Ann. Inst. Statist. Math. **27**(2) (2000), 287–315.

[3] G. C. Cawley and N. L. C. Talbot, *On over-fitting in model selection and subsequent selection bias in performance evaluation*, J. Mach. Learn. Res. **11** (2010), 2079–2107.

[4] I. Csiszar and P. Shields, *Consistency of the BIC order estimator*, Electron. Res. Announc. Amer. Math. Soc. **5** (1999), 123–127.

[5] J. Doob, *Stochastic Processes*, John Wiley & Sons, Inc., New York, 1953

[6] P. A. Jacobs and P. A. W. Lewis, *Discrete time series generated by mixtures. I. Correlational and runs properties*, J. Roy. Statist. Soc. Ser. B **40** (1978), 94–105.
[7] J. Kemeny and J. Snell, *Finite Markov Chains*, Van Nostrand, Princeton, NJ, 1960.
[8] Yu. Kharin, *Robustness in Statistical Forecasting*, Springer, Heidelberg – Dordrecht – New York – London, 2013
[9] Yu. S. Kharin and M. V. Maltsew, *Hypothesis testing for parameters of the Markov chain of conditional order*, Vestsi Nats. Akad. Navuk Belarusi Ser. Fiz.-Mat. Navuk **3** (2012), 5–12. (Russian)
[10] Yu. S. Kharin and M. V. Maltsew, *Markov chain of conditional order: properties and statistical analysis*, Austrian J. Statist. **43**(3) (2014), 205–216.
[11] Yu. S. Kharin and A. I. Petlitskii, *A Markov chain of order s with r partial connections and statistical inference on its parameters*, Discrete Math. Appl. **17**(3) (2007), 295–317.
[12] M. V. Maltsew and Yu. S. Kharin, *On testing of cryptographyc generators output sequences using Markov chains of conditional order*, Informatics **4** (2013), 104–111. (Russian)
[13] A. E. Raftery, *A model for high-order Markov chains*, J. Roy. Statist. Soc. Ser. B **47**(3) (1985), 528–539.
[14] G. Roussas, *Contiguity of Probability Measures: Some Applications in Statistics*, Cambridge University Press, London – New York, 1972.

Belarusian State University, Department of Mathematical Modeling and Data Analysis, Independence Av. 4, Minsk, Belarus
E-mail address: kharin@bsu.by
E-mail address: maltsew@bsu.by