A STATISTICAL TEST BASED ON FREQUENCY STATISTICS OF MARKOV CHAIN WITH PARTIAL CONNECTIONS

Yu.S. Kharin, A.I. Piatlitski
Belarusian State University
Minsk, BELARUS
e-mail: kharin@bsu.by

Abstract

A statistical decision rule for testing of s-dimensional uniformity of discrete time series based on Markov chain with partial connections is constructed. Asymptotic power of the test is found. Numerical results are given.

1 Introduction

Statistical testing of random sequences is one of the most important problems in genetics [4, 7] and other applications. In this paper a new statistical test $T_{MC(s,r)}$ for testing the hypothesis $H_0 = \{\text{observed sequence is an uniformly distributed random sequence (UDRS)}\}$ and the alternative $H_1 = \overline{H_0}$ is constructed; UDRS is a random sequence with independent and uniformly distributed elements. The test $T_{MC(s,r)}$ is based on frequency statistics of new Markovian model [2] – the Markov chain of the s-th order with r partial connections. The power of this test is found and compared with the power of the "overlapping (s + 1)-tuple" test $T$ proposed in [5].

2 Markov chain with partial connections

Let $A = \{0, 1, \ldots, N - 1\}$ be an alphabet with the cardinality $2 \leq N < \infty$; $J_l^i = (j_i, j_{i+1}, \ldots, j_l) \in A^{l-i+1}$ be the multiindex of the $(l - i + 1)$-th order, $l \geq i$; $x_t \in A$ be a homogeneous Markov chain of the s-th order with one-step transition probabilities

$$p_{j_1, \ldots, j_s, j_{s+1}} = P\{x_{t+s} = j_{s+1}|x_{t+s-1} = j_s, \ldots, x_t = j_1\}, \ J_s^{s+1} \in A^{s+1}, \ t \geq 1;$$

$r \in \{1, 2, \ldots, s\}$ be a parameter called the number of connections; $M_r^0 = (m_1^0, \ldots, m_r^0)$ be an integer-valued vector with r ordered components $1 = m_1^0 < m_2^0 < \ldots < m_r^0 \leq s$, called the pattern of connections; $Q = (q_{j_1, \ldots, j_r, j_{r+1}})$, $J_r^{s+1} \in A^{r+1}$ be a stochastic $(r + 1)$-dimensional matrix.

The Markov chain $x_t$ is called [2] the Markov chain of the s-th order with r partial connections $MC(s, r)$ if its one-step transition probabilities have the form:

$$p_{j_1, \ldots, j_s, j_{s+1}} = q_{j_{m_1^0}, \ldots, j_{m_r^0}, j_{s+1}}, \ J_s^{s+1} \in A^{s+1}. \ \ (1)$$

The equation (1) means that the transition probability of the process $x_t$ to the state $j_{s+1}$ depends not on all s preceding states $j_1, j_2, \ldots, j_s$ of the process but only on
r selected states \( j_{m_1}, j_{m_2}, \ldots, j_{m_r} \). Thus, the model (1) is completely determined by 
\( d = N^r(N - 1) + r - 1 \) parameters \( Q \) and \( M_0^r \), instead of \( D = N^s(N - 1) \) parameters 
for Markov chain of the \( s \)-th order. For example, if \( N = 2, s = 32, r = 4 \), then 
\( D > 4 \cdot 10^9, d = 19 \). As it is easy to see, if \( s = r \), then \( MC(s, s) \) is the Markov chain of 
the \( s \)-th order with full connections [1]. The ergodic conditions of \( MC(s, r) \) have been 
determined in [3]. Further, let us assume that the pattern of connections \( M_0^r \) is known 
and \( q_{J_1} > 0, J_{1}^{r+1} \in A^{r+1} \). The Markov chain with partial connections is an ergodic 
process for this case.

Introduce the notation: \( X^n_t = (x_1, x_2, \ldots, x_n) \) is the observed time series of the 
length \( n \); \( \delta_{J_1, K_1} = \prod_{i=1}^{s} \delta_{j_i, k_i} \) is the Kronecker symbol for multiindexes \( J_1 \) and \( K_1 \); 
\( S_t(X_1^n) = (x_{t+m_1-1}^n, \ldots, x_{t+n^2_{m_1}}) \), \( S_t(X_1^n) = (S_t(X_1^n), x_{t+s}) \);

\[
\nu(J_{1}^{r+1}) = \sum_{t=1}^{n-s} \delta_{S_t(X_1^n), J_{1}^{r+1}}, J_{1}^{r+1} \in A^{r+1}, \quad (2)
\]

are the frequency statistics of the Markov chain \( MC(s, r) \); \( \Pi_{K_1}, K_1 \in A^s \), is the stationary 
probability distribution for the ergodic \( MC(s, r) \);

\[
\mu(J_{1}^{r+1}) = P \{ S_t(X_1^n) = J_{1}^{r+1} \} = q_{J_{1}^{r+1}} \sum_{K_1^{s} \in A^s} \delta_{S_t(X_1^n), J_{1}^{r+1}} \Pi_{K_1} \;
\]

\( \hat{\mu}(J_{1}^{r+1}) = \nu(J_{1}^{r+1})/(n-s) \) is the frequency estimator for the probability \( \mu(J_{1}^{r+1}) \).
The estimator \( \hat{\mu}(J_{1}^{r+1}) \) is unbiased and consistent [2]. Let us agree that the point 
instead of any index means summation on all possible values of this index: 
\( \mu(J_{1}^{r}, \cdot) = \sum_{J_{r+1} \in A} \mu(J_{1}^{r+1}) \).

3 Statistical test \( T_{MC(s,r)} \)

Let us construct a statistical test based on (2) for the hypotheses \( H_0: \{ x_t \} \) is UDRS, 
i.e. \( q_{J_{1}^{r+1}} \equiv N^{-1}, J_{1}^{r+1} \in A^{r+1} \); \( H_1: \{ x_t \} \) is the Markov chain \( MC(s, r) \) with

\[
q_{J_{1}^{r+1}} = q_{J_{1}^{r+1}}(n) = N^{-1} \left( 1 + b_{J_{1}^{r+1}}/\sqrt{n-s} \right), \quad J_{1}^{r+1} \in A^{r+1}, \quad (3)
\]

where \( \sum_{J_{r+1} \in A} b_{J_{1}^{r+1}} = 0, \sum_{J_{1}^{r+1} \in A^{r+1}} |b_{J_{1}^{r+1}}| \neq 0 \). The equation (3) means the contiguity 
property of the alternative \( H_1 \): the increase of the length \( n \) implies approaching of 
\( H_1 \) to \( H_0 \) with the rate \( O(n^{-\frac{1}{2}}) \).

Let us define

\[
\xi(J_{1}^{r+1}) = \sqrt{(n-s)N^{r+1}} \left( \hat{\mu}(J_{1}^{r+1}) - N^{-(r+1)} \right), \quad (4)
\]

\[
\rho_{MC(s,r)} = \sum_{J_{1}^{r} \in A^r} \left( \sum_{J_{r+1} \in A} \xi^2(J_{1}^{r+1}) - N^{-1} \xi^2(J_{1}^{r}, \cdot) \right). \quad (5)
\]

253
**Theorem 1.** For the hypothesis $H_0$ the asymptotic probability distribution of the statistic $\rho_{MC(s,r)}$ at $n \to \infty$ is the standard $\chi^2$ distribution with $U = N^r(N - 1)$ degrees of freedom. For the alternative $H_1$ the asymptotic probability distribution of the statistic $\rho_{MC(s,r)}$ at $n \to \infty$ is the noncentral $\chi^2$ distribution with $U$ degrees of freedom and the noncentrality parameter

$$a_{MC(s,r)} = N^{-(r+1)} \sum_{J_1^{r+1} \in A^{r+1}} b_{J_1^{r+1}}^2.$$

This theorem generalizes results of the paper [6].

The statistical test $T_{MC(s,r)}$ for the hypotheses $H_0, H_1$ based on Theorem 1 is:

1) computation of the statistics $\{\nu(J_1^{r+1})\}$ by (2);
2) computation of the statistics $\{\xi(J_1^{r+1}) : J_1^{r+1} \in A^{r+1}\}$ and $\rho_{MC(s,r)}$ according to (4) and (5);
3) computation of the P-value: $P = 1 - G_U(\rho_{MC(s,r)})$, where $G_U(\cdot)$ is the probability distribution function of the standard $\chi^2$ distribution with $U$ degrees of freedom;
4) the decision rule with significance level $\varepsilon \in (0,1)$: if $P > \varepsilon$, then we assign the data to the hypothesis $H_0$, otherwise we take the alternative $H_1$.

**Corollary 1.** If $n \to \infty$, then the power of the test $T_{MC(s,r)}$

$$w \to 1 - G_{U,a_{MC(s,r)}}(G_U^{-1}(1 - \varepsilon)), \quad (6)$$

where $G_{U,a_{MC(s,r)}}(\cdot)$ is the probability distribution function of the noncentral $\chi^2$ distribution with $U$ degrees of freedom and the noncentrality parameter $a_{MC(s,r)}$.

**Corollary 2.** The power of the test $T_{MC(s,r)}$ is greater than the power of the "overlapping $(s + 1)$-tuple" test $T$ proposed in [5].

If parameters $r, M^0_r$ are unknown, then we can modify the test $T_{MC(s,r)}$ using their estimators [3].

**4 Numerical results**

We made numerical experiments to evaluate dependence of the power $w$ on $n$ for the tests $T_{MC(s,r)}, T$. In these experiments the elements $b_{J_1^{r},0}, \ldots, b_{J_1^{r},N-2}$ were generated by the standard generator of the uniformly distributed pseudorandom numbers on the interval (-13,13), while the element $b_{J_1^{r},N-1}$ was equal to $-(b_{J_1^{r},0} + \ldots + b_{J_1^{r},N-2})$. The estimates $\hat{w}$ of the power $w$ were found by the Monte-Carlo method. The dependence of $\hat{w}$ on $n$ at $N = 3$, $s = 8$, $r = 4$, $M^0_r = (1,5,7,8)$ and $a_{MC(s,r)} = 72.6066$ is represented in Figure 1. The dependence for the test $T_{MC(s,r)}$ is plotted at the top of the figure, where the solid line plots the theoretical value (6) of $w$, dotted lines are the upper and the lower 99%-confidence limits, squares indicate the values $\hat{w}$. Similarly, the dependence for the test $T$ is plotted at the bottom of Figure 1. As it is shown in this figure, the power of the test $T_{MC(s,r)}$ is six times larger than the power of the test $T$. Note that the power of the tests doesn’t tend to one at $n \to \infty$ because of contiguous hypotheses.
Figure 1: Dependence of $\hat{w}$ on $n$

References


255