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 problem in genetics [4, 7] and other applications. In this paper a new statistical test $T_{MC(s,r)}$ for testing the hypothesis $H_0 = \{$observed sequence is an uniformly distributed random sequence (UDRS)$\}$ and the alternative $H_1 = \bar{H}_0$ 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 $2 \le N < \infty$; $J_i^l = (j_i, j_{i+1}, \ldots, j_l) \in A^{l-i+1}$ be the multiindex of the $(l-i+1)$-th order, $l \ge i$; $x_t \in A$ be a homogeneous Markov chain of the s-th order with one-step transition probabilities

$$p_{j_1,\ldots,j_s,j_{s+1}} = \mathbf{P}\{x_{t+s} = j_{s+1} | x_{t+s-1} = j_s, \ldots, x_t = j_1\},\ J_1^{s+1} \in A^{s+1},\ t \ge 1;$$

$r \in \{1, 2, \ldots, s\}$ 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 $1 = m_1^0 < m_2^0 < \ldots < m_r^0 \le s$, called the pattern of connections; $Q = (q_{j_1,\ldots,j_r,j_{r+1}})$, $J_1^{r+1} \in A^{r+1}$ 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:

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

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 $j_{m_1^0}, j_{m_2^0}, \ldots, j_{m_r^0}$. Thus, the model (1) is completely determined by $d = N^r(N-1) + r - 1$ parameters Q and M_r^0, 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_r^0 is known and $q_{J_1^{r+1}} > 0$, $J_1^{r+1} \in A^{r+1}$. The Markov chain with partial connections is an ergodic process for this case.

Introduce the notation: $X_1^n = (x_1, x_2, \ldots, x_n)$ is the observed time series of the length n; $\delta_{J_1^s, K_1^s} = \prod_{i=1}^s \delta_{j_i, k_i}$ is the Kronecker symbol for multiindexes J_1^s and K_1^s; $S_t'(X_1^n) = (x_{t+m_1^0-1}, \ldots, x_{t+m_r^0-1})$, $S_t(X_1^n) = (S_t'(X_1^n), x_{t+s})$;

$$\nu\left(J_1^{r+1}\right) = \sum_{t=1}^{n-s} \delta_{S_t(X_1^n), J_1^{r+1}}, \; J_1^{r+1} \in A^{r+1}, \tag{2}$$

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

$$\mu\left(J_1^{r+1}\right) = \mathbf{P}\left\{S_t(X_1^n) = J_1^{r+1}\right\} = q_{J_1^{r+1}} \sum_{K_1^s \in A^s} \delta_{S_1'(K_1^s), J_1^r} \Pi_{K_1^s}^*;$$

$\hat{\mu}(J_1^{r+1}) = \nu(J_1^{r+1})/(n-s)$ 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

$$q_{J_1^{r+1}} = q_{J_1^{r+1}}(n) = N^{-1}\left(1 + b_{J_1^{r+1}}/\sqrt{n-s}\right), \; J_1^{r+1} \in A^{r+1}, \tag{3}$$

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

$$\xi(J_1^{r+1}) = \sqrt{(n-s)N^{r+1}}\left(\hat{\mu}(J_1^{r+1}) - N^{-(r+1)}\right), \tag{4}$$

$$\rho_{MC(s,r)} = \sum_{J_1^r \in A^r} \left(\sum_{j_{r+1} \in A} \xi^2(J_1^{r+1}) - N^{-1}\xi^2(J_1^r, \cdot) \right). \tag{5}$$

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 χ^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 χ^2 distribution with U degrees of freedom and the noncentrality parameter*

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

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 $\left\{\nu(J_1^{r+1})\right\}$ by (2);

2) computation of the statistics $\left\{\xi(J_1^{r+1}) : J_1^{r+1} \in A^{r+1}\right\}$ 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 χ^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 $T_{MC(s,r)}$*

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

where $G_{U,a_{MC(s,r)}}(\cdot)$ is the probability distribution function of the noncentral χ^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_r^0 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 $N = 3$, $s = 8$, $r = 4$, $M_r^0 = (1,5,7,8)$ 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.

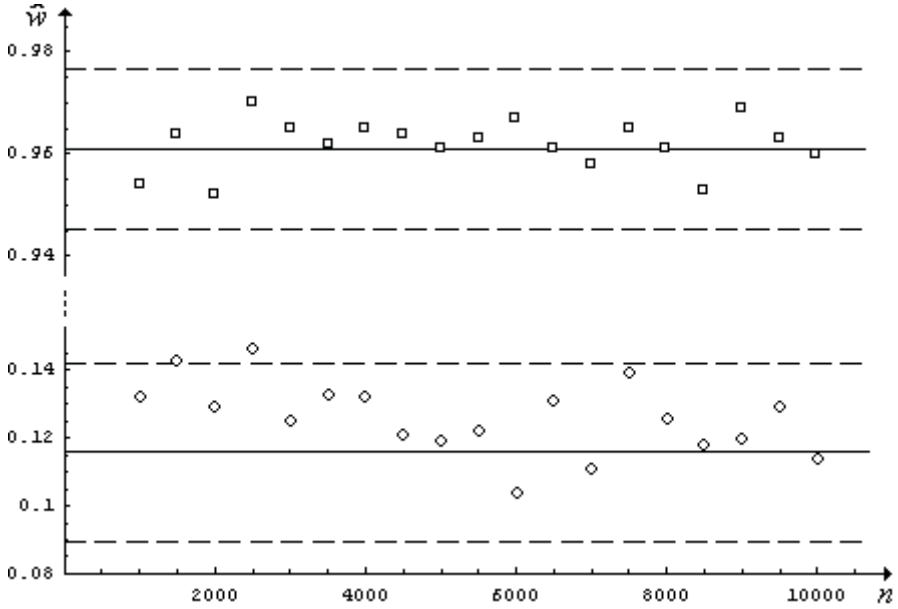

Figure 1: Dependence of $\hat{w}$ on n

References

[1] Doob J.L. (1953). *Stochastic processes*. John Wiley and Sons, New York.

[2] Kharin Yu.S. (2004). Markov chains with r-partial connections and their statistical estimation. *Doklady of National Academy of Siences of Belarus*. Vol. **48**, No. **1**, pp. 40-44. (in Russian).

[3] Kharin Yu.S., Piatlitski A.I. (2006). On estimation of the order of Markov chains with partial connections. *Information systems and technologies*. Part **1**, pp. 156-161. (in Russian).

[4] Knuth D.E. (1997). *The art of computer programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley, Boston.

[5] Marsaglia G.A. (1985). A current view of random number generators. *Computational Science and Statistics: The Interface*. pp. 3-10.

[6] Tikhomirova M.I., Chistyakov V.P. (2003). On two statistics of chi-square type-based on frequencies of tuples of states of a high-order Markov chain. *Discrete Mathematics and Applications* . Vol. **13**, No. **3**, pp. 319-330.

[7] Waterman M.S. (1989). *Mathematical methods for DNA sequences*. CRC Press, Boca Raton.