УДК 519.2: 003.26

**Ю.С. Харин, М.В. Мальцев**

# АЛГОРИТМЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ЦЕПЕЙ МАРКОВА С УСЛОВНОЙ ГЛУБИНОЙ ПАМЯТИ

*Для решения задач компьютерного анализа дискретных временных рядов предлагается новая математическая модель цепи Маркова с условной глубиной памяти. Разработаны алгоритм статистического оценивания параметров модели и алгоритм обнаружения отклонения наблюдаемого временного ряда от модели «чисто случайной» последовательности. Приводятся результаты численных экспериментов.*

## 1. Введение

Задачи компьютерного анализа последовательностей событий или, иначе говоря, дискретных временных рядов часто встречаются в генетике [1], защите информации [2], экономике, медицине и других областях научной и практической деятельности. Одним из важнейших этапов анализа является построение адекватной математической модели, которая бы наиболее близко описывала наблюдаемый процесс или явление.

Для моделирования дискретных временных рядов широко применяется цепь Маркова $s$-го порядка, $s \geq 1$ [3, 4]. Однако число параметров этой модели возрастает экспоненциально при увеличении порядка $s$, что значительно затрудняет ее использование в конкретных приложениях: для статистического оценивания параметров требуется иметь реализацию последовательности далеко не всегда доступной на практике длительности. Поэтому актуальной является задача построения и исследования так называемых малопараметрических моделей [5] цепи Маркова высокого порядка, описываемых меньшим числом параметров, чем полносвязная цепь Маркова порядка $s$.

Примерами таких моделей являются: модель Рафтери [6], в которой требуется лишь один дополнительный параметр для каждого более высокого порядка после первого; цепь Маркова $s$-го порядка с $r$ частичными связями [5], в которой вероятность перехода в текущее состояние $x_t$ зависит не от всех $s$ предыдущих состояний, а лишь от $r$ избранных; цепь Маркова переменного порядка [7]. В этой статье предлагается и исследуется новая малопараметрическая модель дискретных временных рядов – цепь Маркова с условной глубиной памяти.

## 2. Математическая модель

Примем следующие обозначения: $\mathbf{N}$ – множество натуральных чисел, $A = \{0,1,\ldots,N-1\}$ – пространство состояний мощности $N$, $2 \leq N < \infty$; $J_n^m = (j_n, j_{n+1},\ldots, j_{m-1}, j_m) \in A^{m-n+1}$, $m \geq n$, – мультииндекс; $x_t \in A$, $t \in \mathbf{N}$, – однородная цепь Маркова $s$-го порядка $(2 \leq s < +\infty)$, заданная на вероятностном пространстве $(\Omega, F, \mathrm{P})$, с $(s+1)$-мерной матрицей вероятностей одношаговых переходов $P = (p_{j_1^{s+1}})$, $p_{j_1^{s+1}} = \mathrm{P}\{x_{t+s} = j_{s+1} \mid x_{t+s-1} = j_s,\ldots, x_t = j_1\}$, $\forall t \in \mathbf{N}$; $B_* \in \{1,2,\ldots,s-1\}$, $K = N^{B_*} - 1$ – целые числа; $Q^{(1)},\ldots,Q^{(M)}$ – семейство $M$ $(1 \leq M \leq K+1)$ различных квадратных стохастических матриц порядка $N$: $Q^{(m)} = (q_{i,j}^{(m)})$, $i, j \in A$, $1 \leq m \leq M$;

$< J_n^m >= \sum_{k=n}^{m} N^{k-n} j_k \in \{0,1,\ldots, N^{m-n+1} - 1\}$ – числовое представление мультииндекса $J_n^m$;

$\delta_{J_n^m, I_n^m} = \prod_{k=n}^{m} \delta_{j_k, i_k}$ – символ Кронекера для мультииндексов $J_n^m, I_n^m \in A^{m-n+1}$,

$(I_k^l, J_n^m) = K_1^{l-k+m-n+2} \in A^{l-k+m-n+2}$ – конкатенация строк символов $I_k^l$ и $J_n^m$ $(l \geq k,\ m \geq n)$.

Цепь Маркова $s$-го порядка $x_t \in A$ назовем цепью Маркова с условной глубиной памяти, если вероятности одношаговых переходов имеют следующее малопараметрическое представление:

$$p_{J_1^{s+1}} = \sum_{k=0}^{K} \delta_{<J_{s-B_*+1}^{s}>,k} q_{j_{b_k}, j_{s+1}}^{(m_k)} = \begin{cases} q_{j_{b_0}, j_{s+1}}^{(m_0)}, & \text{если } < J_{s-B_*+1}^{s} >= 0, \\ \dots \\ q_{j_{b_K}, j_{s+1}}^{(m_K)}, & \text{если } < J_{s-B_*+1}^{s} >= K, \end{cases} \qquad \textbf{(1)}$$

где $1 \le m_k \le M$, $1 \le b_k \le s - B_*$, $0 \le k \le K$, $\min\limits_{0 \le k \le K} b_k = 1$. Последовательность элементов $J_{s-B_*+1}^{s}$, определяющую условие в формуле (1), назовем базовым фрагментом памяти (БФП) случайной последовательности; $B_*$ – длина БФП. Из (1) видно, что для данной модели состояние $x_t$ процесса в момент времени $t$ зависит не от всех $s$ предыдущих состояний, а от $B_*+1$ состояний $(j_{b_k},\ J_{s-B_*+1}^{s})$, причем численное представление БФП $< J_{s-B_*+1}^{s} >= k$ определяет не только состояние $j_{b_k}$, но и условный порядок цепи Маркова $s_k = s - b_k + 1 \in \{B_*+1, B_*+2, \dots, s\}$, а также матрицу переходов $Q^{(m_k)}$.

Предложенная модель (1) задается следующими параметрами: безусловный порядок $s$ цепи Маркова; длина БФП $B_*$; $K+1$ параметров $\{b_k\}$, определяющих условные порядки цепи Маркова $\{s_k\}$; $K+1$ параметров $\{m_k\}$, определяющих выбор одной из $M$ матриц $Q^{(1)}, \dots, Q^{(M)}$, $M \le K+1$; $M$ стохастических матриц порядка $N$, задаваемых $MN(N-1)$ независимыми параметрами. Таким образом, матрица $P = (p_{J_1^{s+1}})$ вероятностей переходов для цепи Маркова с условной глубиной памяти (1) определяется $D = 2(N^{B_*}+1) + MN(N-1)$ независимыми параметрами.

Заметим, что при $B_* = s-1$, $b_0 = \dots = b_K = 1$, получаем полносвязную цепь Маркова порядка $s$; отметим также, что при $b_0 = \dots = b_K = s - B_*$, получаем полносвязную цепь Маркова порядка $B_*+1$. Если $M = K+1$, то все параметры $\{m_k\}$ различны и каждому БФП соответствует своя матрица вероятностей переходов $Q^{(k)}$. Если $M < K+1$, то среди параметров $\{m_k\}$ имеются совпадающие значения и различным БФП может соответствовать одна и та же матрица $Q^{(k)}$.

### 3. Статистическое оценивание параметров

Вначале найдем условия, при которых цепь Маркова с условной глубиной памяти является эргодической.

**Теорема 1.** Цепь Маркова с условной глубиной памяти является эргодической тогда и только тогда, когда найдется натуральное число $m \in \mathbf{N}$, $s < m < \infty$, такое, что выполняется неравенство:

$$\min_{J_1^s, J_{1+m}^{s+m} \in A^s} \sum_{J_{s+1}^{m} \in A^{m-s}} \prod_{i=1}^{m} \sum_{k=0}^{K} \delta_{<J_{i+s-B_*}^{i+s-1}>,k} q_{j_{b_k+i-1}, j_{i+s}}^{(m_k)} > 0 . \qquad \textbf{(2)}$$

**Доказательство.** Известно [4], что цепь Маркова $s$-го порядка $x_t \in A$, $t \in \mathbf{N}$, является эргодической тогда и только тогда, когда эргодична цепь Маркова первого порядка $X^{(t)} = (x_t, x_{t+1}, \dots, x_{t+s-1})$, $t \in \mathbf{N}$, с расширенным пространством состояний и квадратной матрицей порядка $N^s$ вероятностей одношаговых переходов $\overline{P} = \left(\overline{p}_{<J_1^s>,<J_{s+1}^{2s}>}\right)$, $J_1^s, J_{s+1}^{2s} \in A^s$, где $\overline{p}_{<J_1^s>,<J_{s+1}^{2s}>} = \delta_{J_2^s, J_{s+1}^{2s-1}} p_{J_1^s, j_{2s}}$. Условие эргодичности для цепи Маркова первого порядка $X^{(t)}$ имеет вид [4]: существует натуральное число $m$, $s < m < \infty$, такое что $\min\limits_{J_1^s, J_{1+m}^{s+m} \in A^s} \overline{p}^{(m)}_{<J_1^s>,<J_{1+m}^{s+m}>} > 0$, где $p^{(m)}_{<J_1^s>,<J_{1+m}^{s+m}>}$ – вероятность перехода цепи Маркова $X^{(t)}$ из состояния $J_1^s$ в состояние $J_{1+m}^{s+m}$ за $m$

шагов. Введем обозначения: $I^{(0)} = J_1^s$, $I^{(m)} = J_{1+m}^{s+m}$, $I_-^{(l)} = (i_2^{(l)},\ldots,i_s^{(l)})$, $\bar{I}^{(l)} = (i_1^{(l)},\ldots,i_{s-1}^{(l)})$. Преобразуем $p^{(m)}_{<J_1^s, J_{1+m}^{s+m}>}$ с учетом определения (1):

$$\overline{p}^{(m)}_{<J_1^s, J_{1+m}^{s+m}>} = \sum_{I^{(1)}\in A^s} \overline{p}_{<J_1^s, I^{(1)}>} \overline{p}^{(m-1)}_{<I^{(1)}, J_{1+m}^{s+m}>} = \ldots = \sum_{I^{(1)},\ldots,I^{(m-1)}\in A^s} \prod_{l=0}^{m-1} \overline{p}_{<I^{(l)}, I^{(l+1)}>} = \sum_{I^{(1)},\ldots,I^{(m-1)}\in A^s} \prod_{l=0}^{m-1} \delta_{I_-^{(l)}, \bar{I}^{(l+1)}} p_{I^{(l)}, i_s^{(l+1)}} =$$

$$= \sum_{J_{s+1}^m \in A^{m-s}} \prod_{i=1}^{m} p_{J_i^{i+s-1}, j_{i+s}} = \sum_{J_{s+1}^m \in A^{m-s}} \prod_{i=1}^{m} \sum_{k=0}^{K} \delta_{<J_{i+s-1}^{i+s-1} - B_*>, k} q^{(m_k)}_{j_{b_k+i-1}, j_{i+s}} ,$$

откуда и следует (2). □

В дальнейшем будем рассматривать эргодическую цепь Маркова; ее стационарное распределение обозначим $\pi_{J_1^s} = \mathrm{P}\{x_{t+s-1} = j_s,\ldots,x_t = j_1\}$, $J_1^s \in A^s$, $\forall t \in \mathbf{N}$. Стационарное распределение вероятностей $\Pi = (\pi_{<J_1^s>})$ вычисляется как решение системы линейных алгебраических уравнений [8]: $\overline{P}'\Pi = \Pi$, $\sum_{J_1^s \in A^s} \pi_{J_1^s} = 1$.

Построим оценки максимального правдоподобия (ОМП) для матриц вероятностей переходов $Q^{(1)},\ldots,Q^{(M)}$ по реализации $X_1^n$ длины $n$. Примем обозначения:

$$1 \le l \le s,\ 0 \le l_0 \le s-l,\ A^{s+1}(J_1^l) = \{I_1^{s+1} \in A^{s+1} : I_1^l = J_1^l\},$$
$$A^{1+l_0+l}(j_0^{l_0}, J_1^l) = \{I_1^{1+l_0+l} \in A^{1+l_0+l} : i_1 = j_0,\ I_{2+l_0}^{1+l_0+l} = J_1^l\},$$
$$\nu_{J_1^{s+1}}(n) = \sum_{t=1}^{n-s} \delta_{X_t^{t+s}, J_1^{s+1}}, \quad \nu_{J_1^l}(n) = \sum_{I_1^{s+1}\in A^{s+1}(J_1^l)} \nu_{I_1^{s+1}}(n),\ 1 \le l \le s, \qquad \textbf{(3)}$$
$$\nu^{(l_0)}_{j_0, J_1^l}(n) = \sum_{I_1^{1+l_0+l}\in A^{1+l_0+l}(j_0^{l_0}, J_1^l)} \nu_{I_1^{1+l_0+l}}(n), \quad \xi^{(l_0)}_{j_0, J_1^l}(n) = \frac{\nu^{(l_0)}_{j_0, J_1^l}(n) - n/N^{l+1}}{\sqrt{n/N^{l+1}}}.$$

**Лемма 1.** Для цепи Маркова с условной глубиной памяти $n$-мерное распределение вероятностей при $n > s$ имеет вид:

$$\mathrm{P}\{x_1 = j_1,\ldots,x_n = j_n\} = \pi^0_{J_1^s} \prod_{t=s}^{n-1} \sum_{k=0}^{K} \delta_{<J_{t-B_*+1}^t>, k} q^{(m_k)}_{j_{t-s+b_k}, j_{t+1}},\ j_1,\ldots,j_n \in A, \qquad \textbf{(4)}$$

где $\pi^0_{J_1^s} = \mathrm{P}\{x_1 = j_1,\ldots,x_s = j_s\}$, $J_1^s \in A^s$, – начальное распределение вероятностей цепи Маркова (1).

**Доказательство.** Используя формулу умножения вероятностей, марковское свойство и (1), имеем:

$$\mathrm{P}\{x_1 = j_1,\ldots,x_n = j_n\} = \mathrm{P}\{x_1 = j_1,\ldots,x_s = j_s\} \cdot \mathrm{P}\{x_{s+1} = j_{s+1} \mid x_s = j_s,\ldots,x_1 = j_1\} \times$$
$$\times \mathrm{P}\{x_{s+2} = j_{s+2} \mid x_{s+1} = j_{s+1},\ldots,x_1 = j_1\} \cdot \ldots \cdot \mathrm{P}\{x_n = j_n \mid x_{n-1} = j_{n-1},\ldots,x_1 = j_1\} =$$
$$= \pi^0_{J_1^s} \prod_{t=s}^{n-1} p_{J_{t-s+1}^{t+1}} = \pi^0_{J_1^s} \prod_{t=s}^{n-1} \sum_{k=0}^{K} \delta_{<J_{t-B_*+1}^t>, k} q^{(m_k)}_{j_{t-s+b_k}, j_{t+1}} . \square$$

**Следствие.** Логарифмическая функция правдоподобия для цепи Маркова с условной глубиной памяти имеет вид:

$$l_n(X_1^n, \{Q^{(i)}\}, B_*, \{b_k\}) = \ln \pi^0_{J_1^s} + \sum_{\substack{u,v\in A,\\ w\in A^{B_*}}} \sum_{k=0}^{K} \delta_{<w>,k} \nu^{(l_k)}_{u,wv}(n) \ln q^{(m_k)}_{u,v} ,$$

где $l_k = s - b_k - B_*$.

**Теорема 2.** Если длина БФП $B_*$, значения $\{b_k\}$ и значения $\{m_k = k\}$, $k = 0, 1, \ldots, K$, заданы, то ОМП для вероятностей одношаговых переходов $q_{u,\mathrm{v}}^{(m_k)}$, $1 \le m_k \le K+1$, $u, \mathrm{v} \in A$, имеют вид:

$$\hat{q}_{u,\mathrm{v}}^{(m_k)} = \begin{cases} \sum\limits_{w \in A^{B_*}} \delta_{<w>,k} \dfrac{\nu_{u,w\mathrm{v}}^{(l_k)}(n)}{\nu_{u,w}^{(l_k)}(n)}, \text{ если } \nu_{u,w}^{(l_k)}(n) > 0, \\ 1/N, \text{ если } \nu_{u,w}^{(l_k)}(n) = 0. \end{cases} \qquad \textbf{(5)}$$

**Доказательство.** Для нахождения ОМП требуется решить следующую задачу на условный экстремум:

$$\begin{cases} l_n(X_1^n, \{Q^{(i)}\}_{1 \le i \le M}, B_*, \{b_k\}_{0 \le k \le K}) \to \max\limits_{\{Q^{(i)}\}_{1 \le i \le M}}, \\ \sum\limits_{\mathrm{v} \in A} q_{u,\mathrm{v}}^{(m_k)} = 1, u \in A, 1 \le m_k \le M, w \in A^{B_*}. \end{cases}$$

Эта задача распадается на $N^{B_*+1}$ подзадач отыскания условного максимума:

$$\begin{cases} \sum\limits_{\mathrm{v} \in A} \sum\limits_{k=0}^{K} \delta_{<w>,k} \nu_{u,w\mathrm{v}}^{(l_k)}(n) \ln q_{u,\mathrm{v}}^{(m_k)} \to \max\limits_{q_{u,\mathrm{v}}^{(m_k)}}, \\ \sum\limits_{\mathrm{v} \in A} \sum\limits_{k=0}^{K} \delta_{<w>,k} q_{u,\mathrm{v}}^{(k)} = 1, \end{cases} \quad w \in A^{B_*}, u \in A.$$

Решая указанные задачи методом множителей Лагранжа, приходим к оценкам (5). □

**Замечание.** Если среди параметров $\{m_k\}$, $k = 0, 1, \ldots, K$, имеются одинаковые, т.е. одна и та же матрица переходов соответствует нескольким базовым фрагментам памяти и $M \le K$, то ОМП примут вид:

$$\hat{q}_{u,\mathrm{v}}^{(m_k)} = \begin{cases} \dfrac{\sum\limits_{w \in M_{m_k}} \nu_{u,w\mathrm{v}}^{b_k}}{\sum\limits_{w \in M_{m_k}} \nu_{u,w}^{b_k}}, \text{ если } \sum\limits_{w \in M_{m_k}} \nu_{u,w}^{b_k} > 0, \\ 1/N, \text{ если } \sum\limits_{w \in M_{m_k}} \nu_{u,w}^{b_k} = 0, \end{cases}$$

где $M_i = \{w \in A^{B_*} : m_{<w>} = i\}$, $i = 1, \ldots, M$, $\bigcup\limits_{i=1}^{M} M_i = A^{B_*}$.

Для вычисления ОМП (5) требуется $O(nB_*)$ операций.

**Теорема 3.** Если цепь Маркова с условной глубиной памяти является стационарной, т. е. если выполнено условие эргодичности и начальное распределение совпадает со стационарным, то при $n \to \infty$ оценки (5) являются состоятельными:

$$\hat{q}_{u,\mathrm{v}}^{(m)} \xrightarrow{\mathrm{P}} q_{u,\mathrm{v}}^{(m)}, \ 1 \le m \le M. \qquad \textbf{(6)}$$

**Доказательство.** Согласно [9], для цепи Маркова первого порядка справедлива сходимость по вероятности для частотных оценок к стационарному распределению. По аналогии с доказательством теоремы 1, переходя от цепи Маркова порядка $s$ к цепи Маркова первого порядка с расширенным пространством состояний, имеем сходимость: $\hat{\pi}_{J_1^{s+1}} \xrightarrow{\mathrm{P}} \pi_{J_1^{s+1}}$, где $\hat{\pi}_{J_1^{s+1}} = \nu_{J_1^{s+1}}(n)/(n-s)$, $\pi_{J_1^{s+1}} = \pi_{J_1^s} p_{J_1^{s+1}}$. Выразим частоты $\nu_{u,wv}^{(l_k)}(n)$ и $\nu_{u,w}^{(l_k)}(n)$, входящие в ОМП (5), через частоты $(s+1)$-грамм $\nu_{J_1^{s+1}}(n)$, используя формулы (3):

$$\nu_{u,wv}^{(l_k)}(n) = \sum_{I_1^{B_*+l_k+2} \in A^{B_*+l_k+2}(u^{(l_k)},wv)} \nu_{I_1^{B_*+l_k+2}}(n) = \sum_{I_1^{B_*+l_k+2} \in A^{B_*+l_k+2}(u^{(l_k)},wv)} \sum_{J_1^{s+1} \in A^{s+1}(I_1^{B_*+l_k+2})} \nu_{J_1^{s+1}}(n).$$

Аналогично $\nu_{u,w}^{(l_k)}(n) = \sum_{I_1^{B_*+l_k+1} \in A^{B_*+l_k+1}(u^{(l_k)},w)} \sum_{J_1^{s+1} \in A^{s+1}(I_1^{B_*+l_k+1})} \nu_{J_1^{s+1}}(n)$. Следовательно,

$$\nu_{u,wv}^{(l_k)}(n)/(n-s) \xrightarrow{\mathrm{P}} \sum_{I_1^{B_*+l_k+2} \in A^{B_*+l_k+2}(u^{(l_k)},wv)} \sum_{J_1^{s+1} \in A^{s+1}(I_1^{B_*+l_k+2})} \pi_{J_1^{s+1}} = \mathrm{P}\{x_t = u,\, X_{t+l_k+1}^{t+l_k+B_*+1} = wv\} = \pi_{u,wv}^{(l_k)},$$

$$\nu_{u,w}^{(l_k)}(n)/(n-s) \xrightarrow{\mathrm{P}} \sum_{I_1^{B_*+l_k+1} \in A^{B_*+l_k+1}(u^{(l_k)},w)} \sum_{J_1^{s+1} \in A^{s+1}(I_1^{B_*+l_k+1})} \pi_{J_1^{s+1}} = \mathrm{P}\{x_t = u,\, X_{t+l_k+1}^{t+l_k+B_*} = w\} = \pi_{u,w}^{(l_k)}.$$

Учитывая, что $\pi_{u,wv}^{(l_k)} = \sum_{k=0}^{K} \delta_{<w>,k} \pi_{u,w}^{(l_k)} q_{u,v}^{(m_k)}$, и применяя теорему о функциональном преобразовании сходящихся по вероятности случайных последовательностей [10], получаем (6). □

**Теорема 4.** Если истинные значения $B_*$ и $\{m_k\}$ известны, то ОМП параметров $\{b_k\}$ имеют вид:

$$b_k = \arg\max_{1 \le b \le s-B_*} \sum_{i,j \in A} \nu_{i,wj}^{s-b-B_*}(n) \ln(\hat{q}_{i,wj}^{m_k}),\ k = 1,2,\ldots,K. \qquad \textbf{(7)}$$

**Доказательство**. По аналогии с доказательством теоремы 3, решая задачу на условный экстремум, приходим к оценкам (7). □

Для нахождения оценок (7) требуется $O(nsN^2B_*)$ операций.

Оценки порядка цепи Маркова $s$ и длины БФП $B_*$ находим, решая задачу минимизации информационного функционала Байеса [11]:

$$(\hat{s}, \hat{B}_*) = \arg\min_{2 \le s \le \bar{S},\, 1 \le B \le \bar{B}_*} BIC(s,B),$$

$$BIC(s,B) = -\Big(\sum_{\substack{u,v \in A,\, k=0 \\ w \in A^B}}^{K} \delta_{<w>,k} \nu_{u,wv}^{(s-\hat{b}_k-B)}(n) \ln \hat{q}_{u,v}^{(k)}\Big) + 2N^B \log n, \qquad \textbf{(8)}$$

где $\bar{S} \ge 2$, $1 \le \bar{B}_* \le \bar{S}-1$ – максимально допустимые значения параметров $s$ и $B_*$, оценки $\hat{Q}^{(i)}, i = 1,\ldots,M$, и $\hat{b}_k, k = 0,\ldots,K$, вычисляются по формулам (5) и (7) соответственно.

Для оценивания параметров $\{m_k : k = 0,\ldots,K\}$ можно воспользоваться методом L-средних из кластерного анализа [8], используя в качестве межклассового расстояния евклидову метрику: $\rho(Q^{(i)}, Q^{(j)}) = \| Q^{(i)} - Q^{(j)} \| = \Big(\sum_{k,l \in A} (q_{kl}^{(i)} - q_{kl}^{(j)})^2\Big)^{1/2}$.

Таким образом, алгоритм оценивания параметров цепи Маркова с условной глубиной памяти имеет следующий вид.

**Алгоритм 1:**

1. Задается максимально допустимое значение порядка цепи Маркова $\overline{S}$ и максимально допустимая длина БФП $\overline{B}_*$.

2. По формуле (8) строятся оценки $\hat{S}$, $\hat{B}$.

3. По формуле (7) строятся ОМП $\{\hat{b}_k : k = 0,\ldots,K\}$.

4. По формуле (5) строятся ОМП матриц вероятностей переходов $\hat{Q}^{(1)},\ldots,\hat{Q}^{(K)}$.

5. Используя метод L-средних из кластерного анализа, осуществляется классификация полученных на предыдущем шаге матриц.

Вычислительная сложность алгоритма 1 имеет порядок $O(n\overline{S}\overline{B}_* N^{\overline{B}_*+2})$.

## 4. Статистическое обнаружение отклонения от модели равномерно распределенной случайной последовательности

При построении и оценке надежности систем защиты информации часто возникает задача обнаружения отклонения наблюдаемой последовательности от модели РРСП [2], которая на практике часто называется моделью «чисто случайной» последовательности. Построим тест проверки гипотез: $H_0 = \{x_t \in A$ есть РРСП, т.е. $q_{i,j}^{(m)} = 1/N, \forall i, j \in A, m = 1,2,\ldots,M\}$; $H_1 = \{x_t \in A$ есть цепь Маркова с условной глубиной памяти и вероятностями одношаговых переходов

$$q_{i,j}^{(m)} = q_{i,j}^{(m)}(n) = \frac{1}{N}(1 + \frac{\omega_{i,j}^{(m)}(n)}{\sqrt{n}}) > 0,\ \omega_{i,j}^{(m)}(n) \xrightarrow[n\to\infty]{} \omega_{i,j}^{(m)},\ \sum_{j\in A}\omega_{i.j}^{(m)} = 0,\ \sum_{m=1}^{M}\sum_{i,j\in A}|\,\omega_{i.j}^{(m)}\,| > 0\}.$$

Введем обозначение:

$$\rho(n) = \sum_{\substack{w\in A^{B_*},\\ u,v\in A}}\sum_{k=0}^{K}\delta_{<w>,k}(\xi_{u,wv}^{(l_k)})^2 - \frac{1}{N}\sum_{\substack{w\in A^{B_*},\\ u\in A}}(\sum_{v\in A}\sum_{k=0}^{K}\delta_{<w>,k}\xi_{u,wv}^{(l_k)})^2 . \qquad \textbf{(9)}$$

**Теорема 5**. Если справедлива гипотеза $H_0$, то при $n \to \infty$ распределение вероятностей статистики $\rho(n)$ сходится к стандартному $\chi^2$-распределению с $U = N^{B_*+1}(N-1)$ степенями свободы.

**Доказательство.** Воспользовавшись результатами теоремы 2 и теоремы 3 из [12], и учитывая, что статистика $\rho(n)$ представляет собой квадратичную форму от нормированных частот $\{\xi_i(n)\}$, получаем сформулированный в теореме результат.□

С помощью теоремы 5 построим тест, основанный на статистике $\rho(n)$:

$$\text{принимается } \begin{cases} H_0 : \rho(n) \le \Delta, \\ H_1 : \rho(n) > \Delta, \end{cases} \qquad \textbf{(10)}$$

где $\Delta = G_U^{-1}(1-\alpha)$ – квантиль уровня $1-\alpha$ стандартного $\chi^2$-распределения с $U$ степенями свободы, $\alpha$ – заданный уровень значимости.

Алгоритм обнаружения отклонения от модели РРСП имеет следующий вид.

**Алгоритм 2:**

1. Задается уровень значимости теста $\alpha$.

2. По формулам (3) вычисляются нормированные частоты $\{\xi_{u,wv}^{(l_k)} : u,v \in A,\ w \in B_*,\ k = 0,\ldots,K\}$.

3. По формуле (9) строится статистика $\rho(n)$.

4. На основании (10) принимается решение о том, является ли наблюдаемая последовательность РРСП.

Вычислительная сложность алгоритма 2 имеет порядок $O(nB_*N^{B_*+2})$.

## 5. Численные результаты

Исследуем свойства разработанных алгоритмов методом компьютерного моделирования на модельных и реальных данных.

**Пример 1 (модельные данные).** Пространство состояний $A=\{0,1\}$, $N=2$, $s=4$, $M=2$, $B_*=2$, $b_0=2, m_0=1, b_1=2, m_1=2, b_2=1, m_2=1, b_3=1, m_3=2$. Матрицы $Q^{(1)}$ и $Q^{(2)}$ имеют вид: $Q^{(1)}=\begin{pmatrix}0.18 & 0.82\\ 0.41 & 0.59\end{pmatrix}$, $Q^{(2)}=\begin{pmatrix}0.77 & 0.23\\ 0.09 & 0.91\end{pmatrix}$.

**Случай известных** $s$, $B_*, M$**.** С помощью разработанной программы имитировалась цепь Маркова с условной глубиной памяти с указанными параметрами; согласно алгоритму 1 вычислялись оценки $\hat{Q}^{(1)}$ и $\hat{Q}^{(2)}$ для матриц $Q^{(1)}$ и $Q^{(2)}$ соответственно. Затем вычислялась оценка вариации $\hat{\mathrm{v}}_n^u=\sum_{k=1}^{2}\sum_{i,j=0}^{1}(\hat{q}_{ij}^{(k)}-q_{ij}^{(k)})^2$ для $u$-ой реализации $(u=1,2,\ldots,U)$ длины $n=500,750,1000,\ldots,10000$. Для каждого $n$ оценки $\{\hat{\mathrm{v}}_n^u\}$ вычислялись по $U=1000$ независимым реализациям; затем вычислялось среднее $\hat{\mathrm{v}}_n=\frac{1}{U}\sum_{u=1}^{U}\hat{\mathrm{v}}_n^u$, характеризующее интегральную погрешность построенных оценок (см. рис. 1).

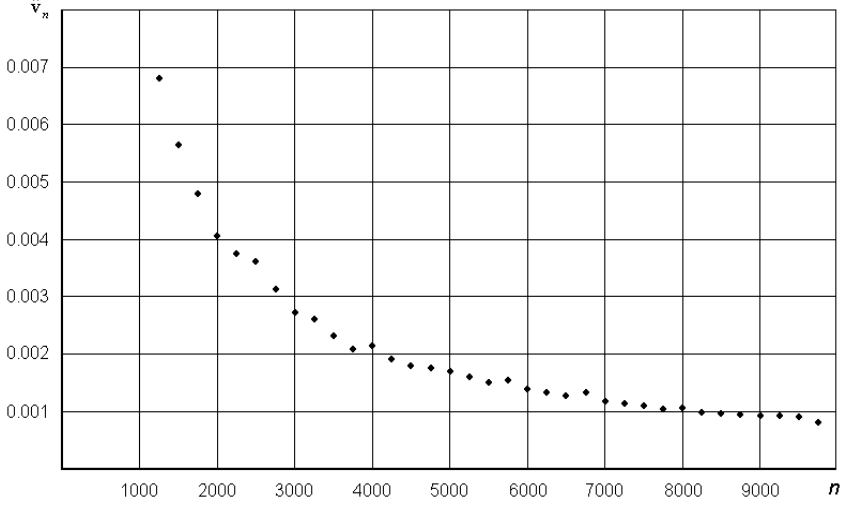


Рис.1: Иллюстрация состоятельности оценок (5)

Аналогичным образом проиллюстрируем состоятельность оценки $\hat{\mathbf{b}}=(\hat{b}_0,\ldots,\hat{b}_K)$, вычисляемой согласно (7). Для этого найдем относительную частоту правильного решения

$$\varepsilon=\frac{1}{U}\sum_{u=1}^{U}\delta_{\hat{\mathrm{b}}^{(u)},\mathrm{b}},\quad \mathbf{b}=(b_0,\ldots,b_K).$$

Таблица 1

| $n$ | 500 | 1000 | 1500 | 2000 | 2500 | 3000 | 3500 | 4000 | $\geq 4250$ |
|---|---|---|---|---|---|---|---|---|---|
| $\varepsilon$ | 0.322 | 0.695 | 0.896 | 0.936 | 0.982 | 0.986 | 0.992 | 0.995 | 1 |

Из таблицы 1 видно, что вероятность правильного решения достаточно быстро стремится к единице при увеличении длины наблюдаемой последовательности.

**Случай неизвестных** $s$, $B_*$**.** С помощью разработанной компьютерной программы имитировалась цепь Маркова с условной глубиной памяти длины $n=50000$ с указанными выше параметрами. Статистическое оценивание осуществлялось согласно алгоритму 1.

1. Максимальное значение порядка $\overline{S}=8$, максимальная длина БФП $\overline{B}_*=4$.

2. В таблицу 2 заносим значения порядка $s$, длины БФП $B$ и соответствующие им значения $BIC(s,B)$. Согласно (8) находим: $\hat{s}=4$, $\hat{B}_*=2$.

Таблица 2

| $(s,B)$ | (2,1) | (3,1) | (3,2) | (4,1) | (4,2) | (4,3) | (5,1) | (5,2) | (5,3) | (5,4) | (6,1) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $BIC(s,B)$ | 31371 | 30116 | 28535 | 27347 | 22852 | 22934 | 27347 | 22852 | 22929 | 23097 | 27347 |
| $(s,B)$ | (6,2) | (6,3) | (6,4) | (7,1) | (7,2) | (7,3) | (7,4) | (8,1) | (8,2) | (8,3) | (8,4) |
| $BIC(s,B)$ | 22852 | 22929 | 23093 | 27347 | 22852 | 22928 | 23092 | 27347 | 22852 | 22928 | 23089 |

Из таблицы 2 видно, что минимум $BIC(s,B)$ достигается при $\hat{B}_*=2$ и при нескольких значениях $s=4,5,\ldots 8$, но фактический порядок модели во всех случаях равен 4, иначе говоря, следующие пары значений $(s,B)$: (4, 2), (5, 2), (6, 2), (7, 2), (8, 2), эквивалентны.

3. Оцениваем параметр $\hat{\mathbf{b}}=(\hat{b}_0,\ldots,\hat{b}_K)$: $b_k=\underset{1\le b\le \hat{s}-\hat{B}_*}{\arg\max}\sum_{i,j\in A}\nu_{i,wj}^{\hat{s}-b-\hat{B}_*}(n)\ln(\hat{q}_{i,wj}^{m_k})$, $k=1,2,\ldots,K$. Получена верная оценка вектора $\hat{\mathbf{b}}=(2,2,1,1)$.

4. Оцениваем четыре матрицы вероятностей переходов (каждому БФП соответствует своя матрица):

$$Q^{(1)}=\begin{pmatrix}0.181 & 0.819\\ 0.400 & 0.600\end{pmatrix},\quad Q^{(2)}=\begin{pmatrix}0.772 & 0.228\\ 0.088 & 0.912\end{pmatrix},\quad Q^{(3)}=\begin{pmatrix}0.183 & 0.817\\ 0.402 & 0.598\end{pmatrix},\quad Q^{(4)}=\begin{pmatrix}0.774 & 0.226\\ 0.089 & 0.911\end{pmatrix}.$$

**Пример 2 (реальные данные).** Исследовалась последовательность длины $n=12500$ для ДНК дрозофилы из генетического банка данных [13]. Пространство состояний $A=\{0,\ 1,\ 2,\ 3\}$, $N=4$, $B_*=2$, порядок $s$ принимал значения 3, 4, …, 11, 16, 24, 32, 48. Для каждого $s$ вычислялась оценка вектора $\mathbf{b}$. Кроме того, вычислялась величина $\Delta_s=\frac{1}{K+1}\sum_{k=0}^{K}\Delta_{sk}$, где $\Delta_{sk}=\sum_{i,j\in A}(\hat{q}_{ij}^{(k)}-1/N)^2$, $K=N^{B_*}-1=15$, характеризующая отклонение наблюдаемой генетической последовательности от РРСП. Для сравнения вычислялась аналогичная оценка $\Delta_s^0$, полученная для РРСП, смоделированной с помощью компьютерной программы. Значения величины $\Delta_s^0$ изменялись в диапазоне от 0.12 до 0.17; значения величины $\Delta_s$ для генетических данных изменялись от 0.46 до 0.48 при различных значениях порядка $s$. Это говорит о том, что исследуемую последовательность ДНК нельзя рассматривать как РРСП. В результате оценивания вектора $\mathbf{b}$ было установлено, что для некоторых БФП значение соответствующей компоненты $b'_k=s-B_*-b_k+1$, $k=0,1,\ldots,K$ не менялось с увеличением порядка $s$ (значение $b'_k=1$ указывает на символ, непосредственно предшествующий базовому фрагменту памяти). Приведем эти БФП и соответствующие им значения компонент $b'_k$: 00 – 1, 01 – 2, 12 – 3, 23 – 1. Проведенные эксперименты показывают, что разработанный в статье алгоритм 2 позволяет выявлять закономерности в генетических последовательностях, используемых в медицинских и биологических исследованиях.

### 6. Заключение

Таким образом, для решения актуальных задач компьютерного анализа дискретных временных рядов в статье предложена новая малопараметрическая модель цепи Маркова $s$-го порядка, ранее в литературе не рассматривавшаяся, – цепь Маркова с условной глубиной памяти. Исследованы вероятностные свойства модели: найдены необходимые и достаточные условия эргодичности, найдено $n$-мерное распределение вероятностей. Построены оценки максимального правдопо-

добия переходных вероятностей, исследованы их асимптотические свойства. Построены оценки максимального правдоподобия параметров $b_0,\ldots,b_K$, характеризующих глубину памяти. Построен тест статистической проверки гипотез для обнаружения отклонения от модели равномерно распределенной случайной последовательности. Проведены компьютерные эксперименты на модельных и на реальных данных, иллюстрирующие работоспособность предложенных в статье алгоритмов.

**Список литературы**


1. Уотермен, М.С. Математические методы для анализа последовательностей ДНК / М. С. Уотермен. – М.: Мир, 1999. – 350 с.

2. Харин, Ю.С. Математические и компьютерные основы криптологии / Ю С. Харин, В И. Берник, Г. В. Матвеев, С. В. Агиевич. – Мн.: Новое знание, 2003. – 381 с.

3. Кемени, Дж. Конечные цепи Маркова / Дж. Кемени, Дж. Снелл. – М.: Наука, 1970. – 272 с.

4. Дуб, Дж. Вероятностные процессы / Дж. Дуб. – М., 1956. – 605 с.

5. Харин, Ю.С. Цепь Маркова с частичными связями ЦМ($s, r$) и статистические выводы о ее параметрах / Ю. С. Харин, А. И. Петлицкий // Дискретная математика. – 2007. – Т. 19, № 2. – С. 109-130.

6. Raftery, A. E. A model for high-order Markov chains / A. E. Raftery // J. Royal Statistical Society. – 1985. – Vol. B-47, № 3. – P. 528–539.

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

8. Харин, Ю. С. Математическая и прикладная статистика / Ю. С. Харин, Е. Е. Жук. – Мн.: БГУ, 2004 – 272 с.

9. Basawa, I.V. Statistical inference for stochastic processes / I. V. Basawa. – AP, 1980. – 435 p.

10. Крамер, Г. Математические методы статистики / Г. Крамер. – М.: Мир, 1975. – 648 с.

11. Csiszar, I. Consistency of the BIC order estimator / I. Csiszar, P. C. Shields // Electronic research announcements of the American mathematical society. – 1999. – Vol. 5. – P. 123-127.

12. Тихомирова, М. И. О двух статистиках типа хи-квадрат, построенных по частотам цепочек состояний сложной цепи Маркова / М. И. Тихомирова, В. П. Чистяков // Дискретная математика. – 2003. – Т. 15, №2. – С. 149 – 159.

13. http://www.ncbi.nlm.nih.gov/Genbank/



*НИИ прикладных проблем математики и информатики*
*Белорусского государственного университета,*
*Минск, пр. Независимости, 4*
*e-mail: kharin@bsu.by, maltsew@mail.ru*



**Yu.S. Kharin, M.V. Maltsew**


## ALGORITHMS FOR STATISTICAL ANALYSIS OF MARKOV CHAIN WITH CONDITIONAL MEMORY DEPTH


A new mathematical model of Markov chain with conditional memory depth is proposed to solve the problems of discrete time series analysis. The algorithm of statistical estimation of the model and the algorithm for detection of deviation of the observed time series from the «purely random» sequence are developed. The results of numerical experiments are presented.