

# ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ СВОЙСТВА ЦЕПЕЙ МАРКОВА ПЕРЕМЕННОГО ПОРЯДКА

М.В. Мальцев

Учреждение Белорусского государственного университета «НИИ прикладных проблем математики и информатики», НИЛ математических методов защиты информации  
Минск, Республика Беларусь  
телефон: +(37529)2785526; e-mail: maltsew@mail.ru

## 1 АННОТАЦИЯ

Рассматривается цепь Маркова переменного порядка (ЦМП). Построены статистические оценки параметров модели, найдены необходимые и достаточные условия эргодичности. Разработан тест на основе частотных статистик ЦМП для выявления зависимости в выходной последовательности криптографического генератора.

Ключевые слова - контекстная функция, равномерно распределенная случайная последовательность, цепь Маркова переменного порядка, частотные статистики.

## 2 ВВЕДЕНИЕ

Важной задачей в защите информации является выявление зависимостей в выходных последовательностях криптографических генераторов [1]. Похожие задачи статистического анализа временных рядов часто встречаются в кибернетике [2], генетике [3], экономике [4], социологии, медицине и во многих других областях научной и практической деятельности. Для моделирования дискретных временных рядов применяются цепи Маркова. Наиболее общей моделью является цепь Маркова  $s$ -го порядка [5]. Однако число параметров  $D = (N-1)N^s$  данной модели возрастает экспоненциально при увеличении порядка. Для статистического оценивания параметров требуется иметь реализацию не всегда доступной на практике длительности. Поэтому построен ряд «малопараметрических» моделей цепи Маркова высокого порядка [6-8], одной из которых является цепь Маркова переменного порядка.

## 3 ЦМП( $s$ ) И ЕЕ ВЕРОЯТНОСТНЫЕ СВОЙСТВА

Пусть  $A = \{0, 1, \dots, N-1\}$  – пространство состояний мощности  $2 \leq N < \infty$ ,  $x_1^k = (x_1, \dots, x_k)$ ,  $x_1^k \in A^k$  –

последовательность символов (строка) из  $k$  элементов,  $x_i^j = (x_i, x_{i+1}, \dots, x_j)$  – фрагмент строки  $x_1^k$  с числом элементов  $|x_i^j| = j - i + 1$ ,  $1 \leq i, j \leq k$ ,  $i \leq j$ ,  $uw = (u_1, u_2, \dots, u_{|u|}, w_1, w_2, \dots, w_{|w|})$  – конкатенация строк  $u, w$ ,  $(X_t \in A)_{t \in \mathbf{Z}}$  – однородная цепь Маркова  $s$ -го порядка, заданная на вероятностном пространстве  $(\Omega, F, P)$ , с матрицей вероятностей одношаговых переходов  $P = (p_{x_1^s, x_{s+1}^s})$ ,

$$p_{x_1^s, x_{s+1}^s} = P\{X_{t+1} = x_{s+1} | X_t = x_s, \dots, X_{t-s+1} = x_1\},$$

где  $x_1^{s+1} \in A^{s+1}$ .

**Определение 1** [6]. Цепь Маркова  $(X_t)_{t \in \mathbf{Z}}$  называется цепью Маркова переменного порядка ЦМП( $s$ ), если её вероятности одношаговых переходов  $p_{x_1^{s+1}}$  имеют вид:

$$p_{x_1^s, x_{s+1}^s} = q_{x_{s-l+1}^s, x_{s+1}^s}, \quad (1)$$

$$0 \leq q_{x_{s-l+1}^s, x_{s+1}^s} \leq 1, \quad l = l(x_1^s), \quad x_1^{s+1} \in A^{s+1}, \quad l \in \{0, 1, \dots, s\},$$

$$l(x_1^s) = \min\{k : P\{X_{t+1} = x_{s+1} | X_t = x_s, \dots, X_{t-s+1} = x_1\} = P\{X_{t+1} = x_{s+1} | X_t = x_s, \dots, X_{t-k+1} = x_{s-k+1}\}\}$$

Соотношение (1) означает, что вероятность перехода в состояние  $x_{s+1}$  зависит не от всех  $s$  предыдущих состояний, а лишь от  $l(x_1^s)$  состояний. Помимо  $l(\cdot)$  в [6] определена контекстная функция  $c(x_1^s) = x_{s-l+1}^s$ , которая цепочке предыдущих состояний ставит в соответствие цепочку из  $l$  значимых состояний – контекст [6]. Если  $l(x_1^s) \equiv s$ , то получаем полностью связную цепь Маркова  $s$ -го порядка; если  $\exists x_1^s \in A^s$ ,  $l(x_1^s) = 0$ , то имеем последовательность независимых случайных величин. Через  $\tau$  обозначим множество значений функции  $c(\cdot)$ .

Функция  $l(\cdot)$  обладает следующим свойством: если  $l(x_1^s) = l_0$ ,  $l_0 \in \{1, 2, \dots, s\}$ , то  $l(y_1^{s-l_0+1} x_{s-l_0+2}^s) \geq l_0$ ,  $\forall y_1^{s-l_0+1} \in A^{s-l_0+1}$ .

**Доказательство.** Предположим, что  $\exists y_1^{s-l_0+1} \in A^{s-l_0+1}$ ,  $l(y_1^{s-l_0+1} x_{s-l_0+2}^s) = l_1 < l_0$ . Из определения контекстной функции имеем:

$$\begin{aligned} P\{X_{s+1} = x_{s+1} | X_s = x_s, \dots, X_{s-l_0+1} = 0\} &= \dots = \\ &= P\{X_{s+1} = x_{s+1} | X_s = x_s, \dots, X_{s-l_0+1} = N-1\}, \end{aligned}$$

что противоречит тому, что  $l(x_1^s) = l_0$ .

Контекстную функцию  $c(\cdot)$  и функцию  $l(\cdot)$  удобно представлять в виде корневого дерева, которое называется контекстным деревом. У каждой вершины в таком дереве может быть не более  $N$  потомков, поскольку каждому узлу (кроме корня) соответствует элемент из пространства состояний  $A$ . Каждому значению контекстной функции соответствует ветвь контекстного дерева. Заметим, что если у каждой вершины контекстного дерева, не являющейся листом, имеется ровно  $N$  потомков, то такое контекстное дерево соответствует полносвязной цепи Маркова  $s$ -го порядка. Такое контекстное дерево называется максимальным контекстным деревом.

**Пример 1.** Пространство состояний  $A = \{0, 1\}$ , порядок  $s = 3$ , контекстная функция  $c(\cdot)$  и соответствующее ей контекстное дерево имеют вид:

$$c(x_1^3) = \begin{cases} 0, x_3 = 0, x_2, x_1 \in A; \\ 0, 1, x_3 = 1, x_2 = 0, x_1 \in A; \\ 0, 1, 1, x_3 = 1, x_2 = 1, x_1 = 0; \\ 1, 1, 1, x_3 = 1, x_2 = 1, x_1 = 1. \end{cases}$$

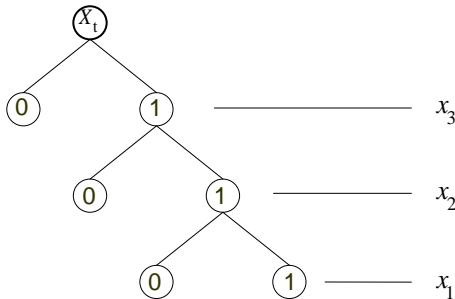


Рис. 1. Контекстное дерево

Найдем условия, при которых ЦМПП ( $s$ ) является эргодической.

**Теорема 1.** Цепь Маркова переменного порядка ЦМПП ( $s$ ) с контекстной функцией  $c(\cdot)$  является эргодической тогда и только тогда, когда найдется такое  $m \geq s$ ,  $m \in \mathbf{N}$ , что

$$\min_{x_1^s, x_{m+1}^{m+s} \in A^s} \sum_{x_{s+1}^m \in A^{m-s}} \prod_{i=1}^m p_{c(x_1^{i+s-1}), x_{i+s}} > 0.$$

**Доказательство.** Переходя от ЦМПП ( $s$ ) к цепи Маркова первого порядка  $X^{(t,s)} = (X_t, \dots, X_{t+s-1})$ ,  $t \in \mathbf{Z}$ ,

с расширенным пространством состояний и используя необходимое и достаточное условие эргодичности для цепи Маркова первого порядка  $X^{(t,s)}$  [9], приходим к требуемому результату.

Обозначим  $\Pi_{x_1^s} = P\{X_1 = x_1, \dots, X_s = x_s\}$ ,  $x_1^s \in A^s$ , – начальное  $s$ -мерное распределение вероятностей ЦМПП ( $s$ ).

**Лемма 1.** Распределение вероятностей реализации  $X = (X_1, \dots, X_n)$  ЦМПП ( $s$ ) имеет вид:

$$P\{X_1 = x_1, \dots, X_n = x_n\} = \Pi_{x_1^s} \cdot \prod_{i=s+1}^n q_{c(x_1^{i-1}), x_i}.$$

**Доказательство.** Используя формулу умножения вероятностей и марковское свойство, приходим к требуемому результату.

#### 4 ОЦЕНИВАНИЕ ПАРАМЕТРОВ МОДЕЛИ

Оценки для переходных вероятностей ЦМПП ( $s$ ), предложенные в [6], имеют вид:

$$\hat{q}_{x_{s-l+1}^s, x_{s+1}} = \frac{v_{x_{s-l+1}^s, x_{s+1}}(n)}{v_{x_{s-l+1}^s}(n)}, \quad (2)$$

где  $v_{x_a^b}(n) = \sum_{i=1}^{n-b+a} \delta_{X_i^{i+b-a}, x_a^b}$  – частотные статистики

ЦМПП ( $s$ ),  $\delta_{x_1^k, y_1^k} = \prod_{i=1}^k \delta_{x_i, y_i}$ ,  $\delta_{x_i, y_i}$  – символ Кронекера.

Покажем, что приведенные оценки являются условными оценками максимального правдоподобия.

**Теорема 2.** Если для реализации  $X = (X_1, \dots, X_n)$  ЦМПП ( $s$ ), определяемой (1), длительности  $n > s$  с известной функцией  $c(\cdot)$  выполнено условие  $v_{x_{s-l+1}^s}(n) > 0$ , то оценки (2) являются условными оценками максимального правдоподобия.

**Доказательство.** Используя результат леммы 1, построим логарифмическую функцию правдоподобия:

$$\begin{aligned} l_n(X, \{q_{\omega, u}\}_{\omega \in \tau, u \in A}) &= \ln \Pi_{x_1^s} + \sum_{i=s+1}^n \ln q_{c(x_1^{i-1}), x_i} = \\ &= \ln \Pi_{x_1^s} + \sum_{\substack{u \in A, \\ \omega \in \tau}} v_{\omega u}(n) \ln q_{\omega, u}. \end{aligned}$$

Экстремальная задача для нахождения оценок максимального правдоподобия имеет вид:

$$\begin{cases} l_n(X, \{q_{\omega, u}\}_{\omega \in \tau, u \in A}) = \ln \Pi_{x_1^s} + \sum_{\substack{u \in A, \\ \omega \in \tau}} v_{\omega u}(n) \ln q_{\omega, u} \rightarrow \max_{\substack{\{q_{\omega, u}\} \\ u \in A, \omega \in \tau}}, \\ \sum_{u \in A} q_{\omega, u} = 1, \omega \in \tau. \end{cases}$$

Используя метод множителей Лагранжа для решения данной задачи, приходим к оценкам (2).

Рассмотрим стационарную цепь Маркова переменного порядка. Тогда оценки (2) являются несмещенными и состоятельными.

Пусть  $(X_t \in A)_{t \in \mathbb{Z}}$  – ЦМПП ( $s$ ), определяемая (1),  $x_1^l, 2 \leq l \leq s$ , – ветвь контекстного дерева. Построим алгоритм оценивания контекстного дерева для ЦМПП ( $s$ ), основанный на проверке следующих вспомогательных гипотез о значимости символов:  $H_0$  – первый символ  $x_1$  в цепочке  $x_1^l$  не является значимым, то есть

$$P\{X_{l+1} = x_{l+1} | X_l = x_l, \dots, X_1 = x_1\} = P\{X_{l+1} = x_{l+1} | X_l = x_l, \dots, X_2 = x_2\}; H_1 - \text{вся цепочка } x_1^l \text{ является значимой. Введем в рассмотрение статистику:}$$

$$\gamma(n) = \sum_{x_1, x_{l+1} \in A} \frac{(v_{x_1^{l+1}}(n) - v_{x_1^l}(n) \hat{p}_{x_1^l, x_{l+1}})^2}{v_{x_1^l}(n) \hat{p}_{x_1^l, x_{l+1}}}.$$

**Теорема 3.** Если справедлива гипотеза  $H_0$ , то при  $n \rightarrow \infty$  распределение статистики  $\gamma(n)$  сходится к  $\chi^2$ -распределению с  $N-1$  степенью свободы.

**Доказательство.** Воспользовавшись тестом для проверки гипотезы о порядке цепи Маркова [10], получаем требуемый результат.

Теорема 3 позволяет построить тест, основанный на статистике  $\gamma(n)$ :

$$\begin{cases} H_0 : \gamma(n) < \Delta, \\ H_1 : \gamma(n) \geq \Delta, \end{cases} \quad (4)$$

где  $\Delta$  – порог, определяемый из заданного уровня значимости  $\alpha$ .

**Следствие 1.** Пусть  $\alpha \in (0, 1)$  и  $\Delta = G_{N-1}^{-1}(1-\alpha)$  – квантиль уровня  $1-\alpha$  стандартного  $\chi^2$ -распределения с  $N-1$  степенью свободы. Тогда при  $n \rightarrow \infty$  размер теста равен  $\alpha$ .

**Доказательство.** Найдем порог  $\Delta$ , при заданном уровне значимости  $\alpha$ :

$$\alpha = P\{H_1 | H_0\} = P\{\hat{\gamma}(n) \geq \Delta | H_0\} = 1 - P\{\hat{\gamma}(n) < \Delta | H_0\} = 1 - G_{N-1}(\Delta), \quad \Delta = G_{N-1}^{-1}(1-\alpha), \text{ откуда и следует требуемый результат.}$$

Численные результаты, полученные в результате компьютерного моделирования, показывают, что алгоритм оценивания контекстного дерева, основанный на проверке вспомогательных гипотез о значимости символов является более точным, чем контекстный алгоритм, предложенный в [6] при малых длинах  $n$  ( $n \leq 10000$ ) реализации ЦМПП.

## 5 ПРОВЕРКА ГИПОТЕЗ О ЗНАЧЕНИИ ПАРАМЕТРОВ ЦМПП

Пусть  $(X_t \in A)_{t \in \mathbb{Z}}$  – ЦМПП ( $s$ ), определяемая (1). Построим тест для проверки гипотез:  $H_0 : (X_t \in A)_{t \in \mathbb{Z}}$  – равномерно распределенная случайная последовательность [1], то есть случайная последовательность, элементы которой независимы в совокупности и имеют равномерное распределение вероятностей  $q_{x_{s-l+1}, x_{s+1}} = 1/N$ ;  $H_1 : (X_t \in A)_{t \in \mathbb{Z}}$  – цепь Маркова переменного порядка с переходными вероятностями одношаговых переходов

$$q_{x_{s-l+1}, x_{s+1}} = q_{x_{s-l+1}, x_{s+1}}(n) = \frac{1}{N} \left( 1 + \frac{\omega_{x_{s-l+1}}(n)}{\sqrt{n}} \right) > 0, \quad \text{где } \omega_{x_{s-l+1}}(n) \xrightarrow{n \rightarrow \infty} \omega_{x_{s-l+1}}, \quad (5)$$

$$\text{причем } \sum_{x_{s+1} \in A} \omega_{x_{s-l+1}} = 0, \quad \sum_{x_{s-l+1} \in A^l} \left| \omega_{x_{s-l+1}} \right| \neq 0.$$

Асимптотическое соотношение (5) означает, что рассматривается континуальное семейство альтернатив.

Введем в рассмотрение следующие случайные величины:

$$\xi_i(n) = \frac{v_i(n) - n/N^{l+1}}{\sqrt{n/N^{l+1}}}, \quad i = i_1^{l+1} \in A^{l+1},$$

$$\rho(n) = \sum_{\substack{k, \\ (i_1, \dots, i_k) \in \tau}} \sum_{i_{k+1}=1}^N \xi^2(i_1, \dots, i_{k+1})(n) - \sum_{\substack{k, \\ (i_1, \dots, i_k) \in \tau}} \left( \sum_{i_{k+1}=1}^N \xi(i_1, \dots, i_{k+1})(n) \right)^2.$$

**Теорема 4.** Если справедлива гипотеза  $H_0$ , то при  $n \rightarrow \infty$  распределение вероятностей статистики  $\rho(n)$  сходится к  $\chi^2$ -распределению с  $M = |\tau|(N-1)$  степенями свободы. Если справедлива гипотеза  $H_1$ , то при  $n \rightarrow \infty$  распределение статистики  $\rho(n)$  сходится к нецентральному  $\chi^2$ -распределению с  $M$  степенями свободы и параметром нецентральности  $a^2$ , определяемому следующей формулой:

$$a^2 = \frac{1}{N|\tau|} \sum_{(x_1, \dots, x_k) \in \tau} \left( \omega_{x_1, \dots, x_k, x_{k+1}} \right)^2.$$

**Доказательство.** Воспользовавшись теоремой 2 из [11] и применив линейное преобразование статистики  $\rho(n)$ , получаем требуемый результат.

С помощью теоремы 4 построим тест, основанный на статистике  $\rho(n)$ :

$$\begin{cases} H_0 : \rho(n) \leq \Delta, \\ H_1 : \rho(n) > \Delta, \end{cases}$$

где  $\Delta$  – порог, определяемый из заданного уровня значимости  $\alpha$ .

**Следствие 2.** Пусть  $\alpha \in (0, 1)$  и  $\Delta = G_{N|\tau}^{-1}(1-\alpha)$  – квантиль уровня  $1-\alpha$  стандартного  $\chi^2$ -распределения с  $U = (N-1)|\tau|$  степенями свободы. Тогда при  $n \rightarrow \infty$  размер теста равен  $\alpha$ .

Доказательство аналогично следствию 1.

**Следствие 3.** Мощность теста  $w$  при  $n \rightarrow \infty$  удовлетворяет следующему асимптотическому соотношению:

$$w \xrightarrow{n \rightarrow \infty} 1 - G_{U,a}(G_U^{-1}(1-\alpha)),$$

где  $G_{U,a}(\cdot)$  – функция нецентрального  $\chi^2$ -распределения с  $U$  степенями свободы и параметром нецентральности  $a$ .

**Доказательство.** Используя определение мощности и результат следствия 2 имеем:

$$\begin{aligned} w &= 1 - P\{H_0|H_1\} = 1 - P\{\rho(n) \leq \Delta|H_1\} \xrightarrow{n \rightarrow \infty} \\ &\xrightarrow{n \rightarrow \infty} 1 - G_{U,a}(\Delta) = 1 - G_{U,a}(G_U^{-1}(1-\alpha)). \end{aligned}$$

Результаты компьютерных экспериментов показывают, что значение мощности построенного теста  $w$  и его оценки превышают соответствующие значения для аналогичного теста из [11], что свидетельствует о более высокой эффективности теста, построенного на основе статистики  $\rho(n)$ .

Отметим, что при увеличении длины реализации ЦМПШ не наблюдается сходимости мощности теста к единице, поскольку рассматривается контигуальное семейство альтернатив, то есть при увеличении длительности  $n$  наблюдаемой последовательности, гипотеза  $H_0$  сближается с гипотезой  $H_1$ :  $H_1 \xrightarrow{n \rightarrow \infty} H_0$ .

## ЛИТЕРАТУРА

[1] Математические и компьютерные основы криптологии / Ю.С. Харин [и др.]. – Минск. : Новое знание, 2003. – 381 с.  
 [2] Медведев, Г.А. Вероятностные методы исследования экстремальных систем / Г.А. Медведев. – М. : Наука, 1967. – 380 с.  
 [3] Уотермен, М.С. Математические методы для анализа последовательностей ДНК / М.С. Уотермен. – М. : Мир, 1999. – 350 с.  
 [4] Ching, W. K. High-order Markov chain models for categorical data sequences / W. K. Ching, E. S. Fung, K. N. Michael // Wiley Periodicals. Inc. Naval Research Logistics. – 2004. – Vol. 51. – P. 557 – 574.

[5] Кемени, Дж. Конечные цепи Маркова / Дж. Кемени, Дж. Снелл. – М. : Наука, 1970. – 272 с.  
 [6] Buhlmann, P. Variable length Markov chains / P. Buhlmann, A. Wyner // The Annals of Statistics. – 1999. – Vol. 27, № 2. – P. 480-513.  
 [7] 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.  
 [8] 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.  
 [9] Дуб, Дж. Вероятностные процессы / Дж. Дуб. – М., 1956. – 605 с.  
 [10] Basawa, I.V. Statistical inference for stochastic processes / I. V. Basawa. – AP, 1980. – 435 p.  
 [11] 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.