УДК 519.2

Статистический анализ частично наблюдаемых выходных последовательностей криптографических генераторов с использованием модели DAR(p)

Для описания частично наблюдаемых выходных последовательностей криптографических генераторов рассмотрена модель дискретной авторегрессии при наличии цензурирования. Найдена логарифмическая функция правдоподобия, и построены приближенные оценки максимального правдоподобия параметров модели.

И.А. БОДЯГИН,
канд. физ.-мат. наук,
заведующий кафедрой математического моделирования и анализа данных

О.В. ДЕРНАКОВА,
магистрант факультета прикладной математики и информатики

УО «Белорусский государственный университет»

Ключевые слова:
модель дискретной авторегрессии, цензурированные наблюдения, метод максимального правдоподобия.

Введение. Неотъемлемыми элементами современных систем криптографической защиты информации являются случайные и псевдослучайные последовательности и их генераторы [1].

Для оценки стойкости криптографических генераторов разработаны различные математические методы криптоанализа, использующие алгебраический аппарат. Однако многие современные криптографические генераторы имеют сложную структуру, что затрудняет применение алгебраических методов для их анализа. В таком случае продуктивным может оказаться моделирование криптографических генераторов с помощью методов теории вероятностей и математической статистики, в частности методов статистического анализа дискретных временны́х рядов [2]. Изучение статистических свойств выходных последовательностей криптографического генератора является важной задачей, поскольку если удастся найти модель, описывающую его выходные последовательности, и оценить параметры такой модели, то появится возможность прогнозировать эти последовательности, что приведет к уязвимости криптографического алгоритма, использующего данный генератор.

Одной из возможных моделей для исследования выходных последовательностей криптографических генераторов является модель DAR(p) [1]. Она позволяет описывать, например, регистры сдвига с линейной обратной связью даже при наличии дополнительного аддитивного зашумления.

На практике данные не всегда наблюдаются или известны полностью. Во многих случаях об истинном значении выходной последовательности генератора в некоторый момент времени известно только, что оно принадлежит некоторому заданному множеству. В этом случае будем говорить, что выходная последовательность наблюдается лишь частично. В математической статистике искажения такого рода называются цензурированием [3, 4].

Теоретический анализ. Рассмотрим $\mathbb{Z}_n$ – кольцо классов вычетов по модулю n, задаваемое множеством целых чисел $\mathbb{Z}_n = \{0,1,...,n-1\}$ и операциями сложения и умножения по модулю n [1].

Пусть временной ряд $X_t \in \mathbb{Z}_n$ описывается дискретной авторегрессионной моделью порядка p (DAR(p)) [1]:

$$X_t = \left(a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + u_t\right) \bmod n, \quad t \geq p, \tag{1}$$

где $a_1, ..., a_p \in \mathbb{Z}_n$ – коэффициенты дискретной авторегрессии, причем $a_p \neq 0$;

$\{u_t\}$ – независимые одинаково распределенные случайные величины с некоторым дискретным распределением вероятностей $\left(u_t \in \mathbb{Z}_n\right)$.

Операция $x \bmod n$ означает нахождение остатка от деления числа x на n. В дальнейшем для сокращения записи $\bmod n$ будем опускать, все операции сложения и умножения будут подразумеваться по модулю n.

Предположим, что

$$P\{u_t = 0\} = q, \quad P\{u_t = i\} = \frac{1-q}{n-1}, \quad q \in [0,1], \quad i \neq 0, \quad i \in \mathbb{Z}_n, \tag{2}$$

причем

$$q > \frac{1-q}{n-1} \Leftrightarrow q > \frac{1}{n}. \tag{3}$$

Условия (2), (3) гарантируют, что u_t принимает значение, равное 0, с вероятностью q чаще, чем любое другое значение, а остальные величины считаются равновероятными.

Также предположим, что известно распределение:

$$P\{X_1 = x_1, ..., X_p = x_p\} = \pi\left(x_1, \ldots, x_p\right), \; x_1, \ldots, x_p \in \mathbb{Z}_n.$$

Наблюдения проводятся в моменты времени $t = \{1, \ldots, T\}$. Будем говорить, что в момент времени t наблюдение цензурировано, если истинное значение временно́го ряда не наблюдается, а известно лишь, что $X_t \in A_t \subseteq \mathbb{Z}_n$, где A_t – наблюдаемое множество, $|A_t| = m_t$.

В данной работе предполагается, что несколько цензурированных наблюдений могут идти подряд, т. е. образовывать серию. Пусть наблюдается s ($s \geq 0$) серий цензурированных значений, имеющих длины k_j, $k_j \geq 1$. Обозначим f_j, l_j первый и последний моменты времени в j-й серии цензурированных значений соответственно.

В каждый момент времени t наблюдается либо истинное значение X_t, либо множество A_t. Необходимо по таким наблюдениям оценить параметры $\theta = (a_1, \ldots, a_p, q)$ модели (1), (2).

Далее в работе будем рассматривать модель дискретной авторегрессии порядка $p = 1$:

$$X_t = \left(aX_{t-1} + u_t\right) \bmod n, \quad t \in \mathbb{Z}, \tag{4}$$

$$P\{X_1 = x_1\} = \pi\left(x_1\right), \quad x_1 \in \mathbb{Z}_n, \tag{5}$$

однако все полученные результаты можно обобщить на случай, когда p принимает значения более высокого порядка.

Для оценивания параметров будем использовать метод максимального правдоподобия.

Построим оценки параметров $\theta = (a, q)$ модели (2), (4) в случае, когда все значения X_t известны точно.

Введем обозначения: $\alpha = \alpha\left(a, X\right) = \sum_{t=2}^{T} \delta_{x_t,\, a x_{t-1}}$, где $a \in \mathbb{Z}_n$, $X = \left(x_1, \ldots, x_T\right)' \in \mathbb{Z}_n^T$,

$\delta_{x_t, a x_{t-1}}$ – δ-символ Кронекера:

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Лемма 1. Пусть имеет место дискретная авторегрессия первого порядка (4). Тогда условную вероятность $P\{X_t = x_t \mid X_{t-1} = x_{t-1}; \theta\}$ можно представить в следующем виде:

$$P\{X_t = x_t \mid X_{t-1} = x_{t-1}; \theta\} = \frac{1-q}{n-1}\left(\frac{q(n-1)}{1-q}\right)^{\delta_{x_t, a x_{t-1}}}. \tag{6}$$

Доказательство. Согласно (2) и (4) имеем:

$$P\{X_t = x_t \mid X_{t-1} = x_{t-1}; \theta\} =$$
$$= P\{aX_{t-1} + u_t = x_t \mid X_{t-1} = x_{t-1}; \theta\} =$$
$$= P\{u_t = x_t - a x_{t-1}; \theta\} = \begin{cases} q, & x_t - a x_{t-1} = 0, \\ \dfrac{1-q}{n-1}, & x_t - a x_{t-1} \neq 0. \end{cases}$$

Запишем последнее равенство через δ-символ Кронекера:

$$P\{X_t = x_t \mid X_{t-1} = x_{t-1}; \theta\} =$$
$$= q^{\delta_{x_t, a x_{t-1}}} \left(\frac{1-q}{n-1}\right)^{1-\delta_{x_t, a x_{t-1}}} = \frac{1-q}{n-1}\left(\frac{q(n-1)}{1-q}\right)^{\delta_{x_t, a x_{t-1}}}.$$

Теорема 1. Если имеет место дискретная авторегрессионная модель первого порядка (4) и все значения X_t известны точно, то логарифмическая функция правдоподобия может быть вычислена следующим образом:

$$l(\theta;X)=(T-1)\ln\frac{1-q}{n-1}+\alpha\ln\frac{q(n-1)}{1-q}.$$

Доказательство. Функция правдоподобия по определению равна [5]:

$$L(\theta;X)=P\{X_1=x_1,\ldots,X_T=x_T;\theta\}=$$
$$=P\{X_1=x_1\}\prod_{t=2}^{T}P\{X_t=x_t\mid X_{t-1}=x_{t-1};\theta\},$$

но, поскольку начальное распределение (5) не зависит от оцениваемых параметров, мы можем записать:

$$L(\theta;X)=\prod_{t=2}^{T}P\{X_t=x_t\mid X_{t-1}=x_{t-1};\theta\}.$$

Представим $P\{X_t=x_t\mid X_{t-1}=x_{t-1};\theta\}$ в виде (6), тогда

$$L(\theta;X)=\left(\frac{1-q}{n-1}\right)^{T-1}\prod_{t=2}^{T}\left(\frac{q(n-1)}{1-q}\right)^{\delta_{x_t,ax_{t-1}}}.$$

$$l(\theta;X)=\ln L(\theta;X)=(T-1)\ln\frac{1-q}{n-1}+\alpha\ln\frac{q(n-1)}{1-q}.$$

Рассмотрим случай одной серии цензурированных наблюдений.

Теорема 2. Пусть имеет место дискретная авторегрессионная модель первого порядка (4), значения $X_1,\ldots,X_{T-k-1}$ и X_T известны точно, а вместо $X_{T-k},\ldots,X_{T-1}$ наблюдаются случайные события $\{X_t\in A_t\}$, $A_t=\{b_1,\ldots,b_{m_t}\}$, $m_t\le n$, где $t=\overline{T-k,\,T-1}$, $k=\overline{1,\,T-2}$. Тогда логарифмическая функция правдоподобия имеет следующий вид:

$$l(\theta;X)=(T-1)\ln\frac{1-q}{n-1}+\alpha\ln\frac{q(n-1)}{1-q}+\ln\left(\sum_{i_1=1}^{m_{T-k}}\cdots\sum_{i_k=1}^{m_{T-1}}\left(\frac{q(n-1)}{1-q}\right)^{\Delta_{i_1,\ldots,i_k}}\right),$$

где $\alpha=\alpha(a,X)=\sum_{t=2}^{T-k-1}\delta_{x_t,ax_{t-1}}$, $\Delta_{i_1,\ldots,i_k}=\delta_{ax_{T-k-1},b_{i_1}}+\sum_{j=1}^{k-1}\delta_{ab_{i_j},b_{i_{j+1}}}+\delta_{ab_{i_k},x_T}$, $i_1=\overline{1,m_{T-k}}$, $\ldots$, $i_k=\overline{1,m_{T-1}}$.

Доказательство. Преобразуем функцию правдоподобия:

$$L(\theta;X)=P\{X_1=x_1,\ldots,X_T=x_T;\theta\}=P\{X_1=x_1,\ldots,X_{T-k-2}=x_{T-k-2}\mid X_{T-k-1}=x_{T-k-1};\theta\}\times$$
$$\times P\{X_{T-k-1}=x_{T-k-1},X_{T-k}\in A_{T-k},\ldots,X_{T-1}\in A_{T-1},X_T=x_T;\theta\}=P\{X_1=x_1,\ldots,X_{T-k-1}=x_{T-k-1};\theta\}\times$$
$$\times\sum_{i_1=1}^{m_{T-k}}\cdots\sum_{i_k=1}^{m_{T-1}}P\{X_{T-k}=b_{i_1},\ldots,X_{T-1}=b_{i_k},X_T=x_T\mid X_{T-k-1}=x_{T-k-1};\theta\}.$$

Запишем вероятность, стоящую под знаком суммы, в виде произведения и представим каждый из множителей в виде (6). Получим:

$$L(\theta;X)=L(\theta;x_1,\ldots,x_{T-k-1})\sum_{i_1=1}^{m_{T-k}}\cdots\sum_{i_k=1}^{m_{T-1}}\left(P\{X_{T-k}=b_{i_1}\mid X_{T-k-1}=x_{T-k-1};\theta\}\times\right.$$
$$\left.\times\prod_{j=1}^{k-1}P\{X_{T-k+j}=b_{i_{j+1}}\mid X_{T-k+j-1}=b_{i_j};\theta\}P\{X_T=x_T\mid X_{T-1}=b_{i_k};\theta\}\right)=$$
$$=L(\theta;x_1,\ldots,x_{T-k-1})\left(\frac{1-q}{n-1}\right)^{k+1}\sum_{i_1=1}^{m_{T-k}}\cdots\sum_{i_k=1}^{m_{T-1}}\left(\frac{q(n-1)}{1-q}\right)^{\Delta_{i_1,\ldots,i_k}}.$$

Тогда логарифмическая функция правдоподобия будет иметь следующий вид:

$$l(\theta;X)=l(\theta;x_1,\ldots,x_{T-k-1})+(k+1)\ln\frac{1-q}{n-1}+\ln\sum_{i_1=1}^{m_{T-k}}\cdots\sum_{i_k=1}^{m_{T-1}}\left(\frac{q(n-1)}{1-q}\right)^{\Delta_{i_1,\ldots,i_k}}=$$
$$=(T-1)\ln\frac{1-q}{n-1}+\alpha\ln\frac{q(n-1)}{1-q}+\ln\sum_{i_1=1}^{m_{T-k}}\cdots\sum_{i_k=1}^{m_{T-1}}\left(\frac{q(n-1)}{1-q}\right)^{\Delta_{i_1,\ldots,i_k}}$$

Обобщим результат теоремы 2 на случай, когда имеется s серий цензурированных наблюдений.

Теорема 3. Пусть T_o – множество моментов времени, когда значения X_t известны точно. Пусть имеет место дискретная авторегрессионная модель первого порядка (4) и наблюдается s серий цензурированных значений, имеющих длины k_s, $s \geq 1$, $k_s \geq 1$. Тогда логарифмическая функция правдоподобия имеет следующий вид:

$$l(\theta; X) = (T-1)\ln\frac{1-q}{n-1} + \alpha \ln\frac{q(n-1)}{1-q} + \\ + \sum_{j=1}^{s} \ln\left(\sum_{i_1=1}^{m_{f_j}} \cdots \sum_{i_{k_j}=1}^{m_{l_j}} \left(\frac{q(n-1)}{1-q} \right)^{\Delta_{i_1,\ldots,i_{k_j}}} \right),$$

где $\alpha = \alpha(a, X) = \sum\limits_{\substack{t=2,\\ t,\, t-1 \in T_o}}^{T} \delta_{x_t,\, a x_{t-1}}$,

$$\Delta_{i_1,\ldots,i_{k_j}} = \delta_{a x_{f_j-1}, b_{i_1}} + \sum_{r=1}^{k_j-1} \delta_{a b_{i_r}, b_{i_{r+1}}} + \delta_{a b_{i_{k_j}}, x_{l_j+1}},$$

$i_1 = \overline{1, m_{f_j}}$, …, $i_{k_j} = \overline{1, m_{l_j}}$, $j = \overline{1, s}$.

Доказательство. Аналогично доказательству теоремы 2.

Алгоритм оценивания. Таким образом, получена логарифмическая функция правдоподобия, максимизируя которую, можно найти оценки максимального правдоподобия параметров модели.

В случае когда все значения временно́го ряда X_t известны точно, оценки параметров модели можно находить следующим образом. Построим матрицу $(\nu_{ij})_{i,\,j=1}^{n}$ частот переходов из состояния x_{t-1}, равного i, в состояние x_t, равное j:

$$\nu_{ij} = \sum_{t=2}^{T} \delta_{x_{t-1}, i} \delta_{x_t, j}, \quad i, j = \overline{0, n-1}.$$

Тогда, максимизируя величину $\alpha(a, X) = \sum\limits_{i=0}^{n-1} \nu_{i, ai}$, находим оценку максимального правдоподобия параметра a:

$$\hat{a} = \arg\max_a \alpha(a, X).$$

Построить оценку максимального правдоподобия параметра q можно по следующей формуле:

$$\hat{q} = \frac{\alpha(\hat{a}, X)}{T-1}.$$

При наличии цензурирования получить оценки в явном виде не удалось, т. к. функция правдоподобия имеет сложный нелинейный вид.

Для построения приближенных оценок $\hat{a}$ и $\hat{q}$ использовали следующий алгоритм.

При $i = 0, \ldots, n-1$:

1. При $a = i$ находим оценку $\hat{q}^{(i)}$ сеточным методом на интервале [0; 1] с заданной точностью ε.
2. Вычисляем значение логарифмической функции правдоподобия при $a = i$ и $q = \hat{q}^{(i)}$.

В качестве оценок $(\hat{a}, \hat{q})$ выбираем ту пару $(i, \hat{q}^{(i)})$, при которой значение функции максимально.

Экспериментальная часть. Был проведен компьютерный эксперимент на модельных данных для эмпирического оценивания 1) вариации оценки параметра q; 2) вероятности ошибочного нахождения параметра a. Эмпирическое оценивание проводилось по методу Монте-Карло.

Эксперимент состоял в следующем: моделировался N раз временной ряд, описываемый дискретной авторегрессионной моделью первого порядка длины T с параметрами n, a и q; если $X_t \in A_t$, то с вероятностью γ значение X_t считалось цензурированным; по сгенерированным данным оценивались параметры $\hat{a}^{(i)}$ и $\hat{q}^{(i)}$.

Вариацию оценки $\hat{q}$ и вероятность $\hat{P}\{\hat{a} \neq a\}$ находим следующим образом:

$$V\{\hat{q}\} = E\{(\hat{q} - q)^2\} = \frac{1}{N}\sum_{i=1}^{N}(\hat{q}^{(i)} - q)^2,$$

$$\hat{P}\{\hat{a} \neq a\} = 1 - \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{a}^{(i)}, a}.$$

Результаты и их обсуждение. Начальные значения параметров:

$n = 16$, $a = 11$, $q = 0{,}7$, $\gamma = 0{,}7$,

$A_t = A = \{0, 1, 4, 5\}$, $N = 1000$.

T изменяется от 10 до 150 с шагом 10 и от 150 до 1000 с шагом 50.

Содержательный смысл рассматриваемых экспериментов заключается в следующем. Допустим, устройство генерирует четыре бита за такт. Каждая следующая итерация линейно зависит от предыдущей с добавлением аддитивного искажения. В силу

особенностей устройства съема, если второй и четвертый биты равны нулю, то с вероятностью γ первый и третий биты не снимаются.

Результаты проведенных экспериментов представлены на рис. 1, 2.

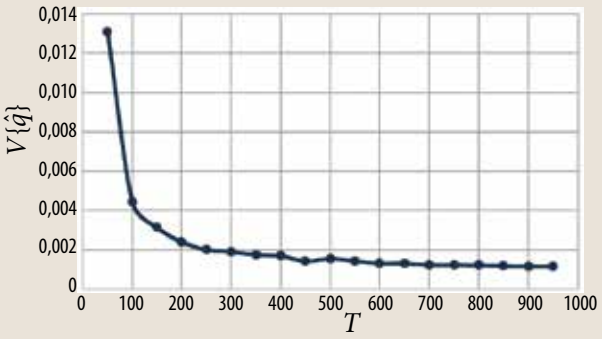

Рисунок 1 – График зависимости вариации оценки $\hat{q}$ от длины T временно́го ряда

Так как значение вариации оценки $\hat{q}$ уменьшается с ростом длины временно́го ряда T, предполагается, что имеет место сходимость оценки $\hat{q}$ к истинному значению параметра q в среднем квадратичном.

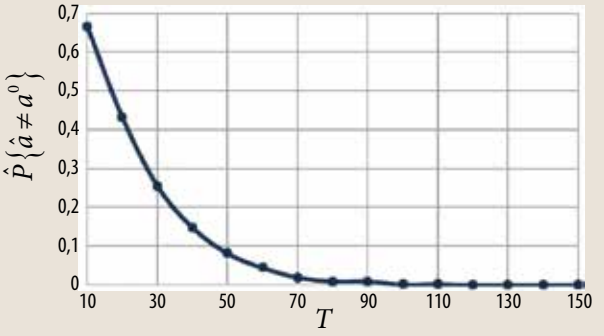

Рисунок 2 – График зависимости вероятности $\hat{P}\{\hat{a} \neq a^0\}$ от длины T временно́го ряда

Поскольку с ростом T вероятность того, что $\hat{a} \neq a^0$, уменьшается, можно допустить, что $\hat{a}$ сходится к a^0 по вероятности.

Заключение. Таким образом, в настоящей работе получены следующие основные результаты:

1. Найдены оценки максимального правдоподобия параметров модели дискретной авторегрессии первого порядка в случае полных данных.
2. Найдена функция правдоподобия в случае наличия цензурированных наблюдений, предложен алгоритм построения приближенных оценок максимального правдоподобия для параметров модели дискретной авторегрессии первого порядка.
3. Проведены компьютерные эксперименты. Практические результаты позволяют предположить, что имеет место сходимость оценок по вероятности.

ЛИТЕРАТУРА

1. **Харин, Ю.С.** Криптология: учебник / Ю.С. Харин [и др.]. – Минск: БГУ, 2013. – 511 с. – (Классическое университетское издание).
2. **Мальцев, М.В.** Моделирование и распознавание криптографических генераторов на основе цепей Маркова условного порядка: автореф. дис. канд. физ.-мат. наук: 05.13.19 / М.В. Мальцев; Белорус. гос. ун-т. – Минск, 2015. – 24 с.
3. **Park, J.W.** Censored time series analysis with autoregressive moving average models. / J.W. Park, M.G. Genton, S.K. Ghosh // The Canadian journal of statistics. – 2007. – Vol. 35, No. 1. – P. 151–168.
4. **Zhidong, B.** Statistical analysis for rounded data / B. Zhidong, Z. Shurong, Z. Baozue, Hu Guorong // Journal of Statistical Planning and Inference. – 2009. – Vol. 139, No. 8. – P. 2526–2542.
5. **Харин, Ю.С.** Теория вероятностей, математическая и прикладная статистика: учебник / Ю.С. Харин, Н.М. Зуев, Е.Е. Жук. – Минск: БГУ, 2011. – 463 с. – (Классическое университетское издание).

The censored discrete time series autoregression model is considered for describing the partially observed output sequences of cryptographic generators. The log-likelihood function is found and approximate maximum likelihood estimators of the model parameters are constructed.

Получено 24.11.2017.