# Марковские свойства преобразования «Book Stack»

Бзикадзе А.В., студент кафедры статистического моделирования СПбГУ, seryrzu@gmail.com
Некруткин В.В., кандидат физико-математических наук, доцент кафедры статистического моделирования СПбГУ, vnekr@statmod.ru

## Аннотация

В докладе изучаются вероятностные свойства преобразования «Book Stack», предложенного Б. Я. Рябко (Пробл. передачи. инф., т. 16, вып. 4, 1980) в качестве процедуры сжатия информации, в случае, когда «входная» последовательность процедуры представляет собой однородную цепь Маркова. Особое внимание уделяется ситуации, когда «входная» последовательность является последовательностью независимых одинаково распределенных случайных величин. Показано, что предложенный Б.Я. Рябко и А.И. Пестуновым (Пробл. передачи. инф., т. 40, вып. 1, 2004) тест для проверки гипотезы о том, что «входная» повторная выборка соответствует дискретному равномерному распределению с известным носителем, и основанный на том, что соответствующий критерий применяется в «выходной» последовательности процедуры, будет иметь, вообще говоря, меньшую мощность, чем тот же критерий, примененный ко «входной» последовательности.

## Введение

В статье [1] предложено преобразование, названное Book Stack (в дальнейшем, BS-преобразование) и используемое в качестве простой и наглядной процедуры сжатия информации. В англоязычной литературе (например, [2]) более распространено название Move-to-Front.

Дадим формальное описание BS-преобразования. Пусть $\mathbb{S} = \{1, 2, \ldots, S\}$ и $\mathfrak{S}_S$ — множество всевозможных перестановок чисел из $\mathbb{S}$. Для любого $x \in \mathbb{S}$ и любой перестановки $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_S)^{\mathrm{T}} \in \mathfrak{S}_S$ обозначим $i_0 = i_0(\alpha, x)$ такой индекс, что $\alpha_{i_0} = x$. Тогда

$$f(\alpha, x)[i] \stackrel{\text{def}}{=} \begin{cases} x & \text{при } i = 1, \\ \alpha_i & \text{при } i > i_0, \\ \alpha_{i-1} & \text{при } 1 < i \leqslant i_0. \end{cases}$$

Тем самым мы получили отображение

$$f = \Big(f(\alpha, x)[1], \ldots, f(\alpha, x)[S]\Big)^{\mathrm{T}} : \mathfrak{S}_S \times \mathbb{S} \mapsto \mathfrak{S}_S,$$

которое и называется BS-преобразованием. Рассмотрим последовательность случайных величин $\{\eta_i\}_{i\geqslant 1}$, предполагая, что $\eta_i \in \mathbb{S}$ при любом $i$.

Введем последовательность векторов $\{\Xi_n \in \mathfrak{S}_S\}_{n\geqslant 0}$ так, что для

$$\Xi_i = (\Xi_i[1], \Xi_i[2], \ldots, \Xi_i[S])^{\mathrm{T}} = f(\Xi_{i-1}, \eta_i), \tag{1}$$

при $i \geqslant 1$, а $\Xi_0$ — вообще говоря, случайный вектор, принимающий значения во множестве перестановок $\mathfrak{S}_S$. Предполагается, что этот случайный вектор $\Xi_0$ не зависит от $\{\eta_i\}_{i\geqslant 1}$.

Наконец, определим последовательность $\{\xi_i\}_{i\geqslant 1}$, где $\xi_i \in \mathbb{S}$ задается как решение уравнения $\eta_i = \Xi_{i-1}[\xi_i]$. Заметим, что для любого $i \geqslant 1$ это решение существует и единственно, так как $\Xi_{i-1}$ является некоторой перестановкой чисел $1, 2, \ldots, S$, а $\eta_i \in \mathbb{S}$.

Результаты, анонсируемые в докладе, можно разделить на 3 части. В первой части (Предложение 1) показано, что в случае, если случайные величины $\{\eta_i\}_{i\geqslant 1}$ образуют однородную марковскую цепь (далее — ОМЦ), последовательность случайных величин $\{\Xi_i\}_{i\geqslant 1}$ также образует ОМЦ. При этом из эргодичности ОМЦ $\{\eta_i\}_{i\geqslant 1}$ следует, что у $\{\Xi_i\}_{i\geqslant 1}$ есть ровно один непериодический эргодический класс и, быть может, несколько несущественных состояний.

Далее подробно изучается частный случай, когда $\{\eta_i\}_{i\geqslant 1}$ независимы и одинаково распределены. Тогда (Теорема 1) оказывается, что, последовательность $\xi_i$ сходится к некоторому предельному распределению, а для частот последовательности $\xi_i$ выполняется вариант закона больших чисел.

Наконец, этот последний результат используется для исследования статистического Book Stack-критерия (и его вариантов), предложенного в [3]. Нетрудно показать, что случайные величины $\{\xi_i\}_{i\geqslant 1}$ являются независимыми и равномерно распределенными на множестве $\mathbb{S}$ (последнее будет обозначаться как $\xi_i \in \mathrm{U}_S$) тогда и только тогда, когда последовательность $\{\eta_i\}_{i\geqslant 1}$ обладает таким же свойством.

В статье [3] предложен статистический критерий для проверки гипотезы $\mathrm{H}_0$ о том, что повторная независимая выборка $\eta_i$ взята из распределения $\mathrm{U}_S$. Общую идею статистических тестов, основанных на BS-преобразовании, можно описать так: эта гипотеза проверяется с помощью случайных величин $\xi_i$, а не исходных $\eta_i$.

Для проверки гипотезы $\mathrm{H}_0$ существует много статистических критериев (см., например, [4]), среди которых наиболее популярными являются критерий $\chi^2$ и критерий отношения правдоподобия. В настоящей работе показано,

что в условиях альтернативы $\mathcal{P} \stackrel{\text{def}}{=} \mathcal{L}(\eta_i) \neq \mathrm{U}_S$ (здесь и далее $\mathcal{L}(\delta)$ обозначает распределение случайной величины $\delta$) критерии отношения правдоподобия и $\chi^2$, примененные к последовательности $\{\xi_i\}_{i\geqslant 1}$, будут при больших объемах выборки (и при некоторых дополнительных условиях) менее мощными, чем такие же критерии, примененные к исходной последовательности $\{\eta_i\}_{i\geqslant 1}$. Аналогичный факт оказывается верным и для нескольких других критериев.

Доказательство этих утверждения основано на результате Теоремы 2, показывающем, что предельное распределение последовательности $\xi_i$ оказывается «ближе» к равномерному $\mathrm{U}_S$, чем $\mathcal{P} \neq \mathrm{U}_S$.

Введем дополнительное обозначение. А именно, для любого $\alpha \in \mathfrak{S}_S$ определим

$$\mathrm{C}_\alpha^{\mathfrak{S}_S} = \{\beta \mid \text{существует } k : f(\beta, k) = \alpha\} \subset \mathfrak{S}_S. \tag{2}$$

## Результаты

### *Марковское свойство последовательности* $\Xi_i$

В этом разделе предполагается, что выполнены следующие условия:

a) последовательность $\{\eta_n\}_{n\geqslant 1}$ является ОМЦ с фазовым пространством $\mathbb{S}$, переходной матрицей $\mathbf{P}^{(\eta)} = (p_{ij})$ и начальным распределением $\left(p_1^{(1)}, p_2^{(1)}, \ldots, p_S^{(1)}\right)$,

b) марковская цепь $\{\eta_n\}_{n\geqslant 1}$ и случайный вектор $\Xi_0 \in \mathfrak{S}_S$, имеющий распределение $(\pi_\beta^{(0)}, \beta \in \mathfrak{S}_S)$, независимы.

**Предложение 1** *1. Последовательность* (1) *образует ОМЦ с фазовым пространством* $\mathfrak{S}_S$*, начальным распределением*

$$\mathrm{P}(\Xi_1 = \alpha) = p_{\alpha[1]}^{(1)} \sum_{\beta \in \mathrm{C}_\alpha^{\mathfrak{S}_S}} \pi_\beta^{(0)}, \quad \alpha \in \mathfrak{S}_S,$$

*и матрицей переходных вероятностей* $\mathbf{P}^{(\Xi)} = \left(p_{\alpha\beta}^{(\Xi)}\right)$

$$p_{\alpha\beta}^{(\Xi)} = \begin{cases} p_{\alpha[1]\beta[1]} & \textit{при } \alpha \in \mathrm{C}_\beta^{\mathfrak{S}_S}, \\ 0 & \textit{иначе}, \end{cases}$$

*где* $\alpha, \beta \in \mathfrak{S}_S$*, а множество* $\mathrm{C}_\alpha^{\mathfrak{S}_S}$ *введено в* (2)*.*

*2. Если дополнительно потребовать, чтобы входная ОМЦ $\{\eta_i\}_{i=1}^{\infty}$ была эргодической, то марковская цепь $\{\Xi_n\}_{n\geqslant 1}$ будет иметь ровно один непериодический эргодический класс и, быть может, несколько несущественных состояний. Если же $p_{ij} > 0$ при всех $i, j$, то несущественных состояний нет.*

**Замечание 1** *В условиях второго пункта Предложения 1 у марковской цепи $\{\Xi_n\}_{n\geqslant 1}$ имеется стационарное распределение $\Pi_S = (\pi_\alpha,\ \alpha \in \mathfrak{S}_S)$, причем $\pi_\alpha = 0$, если $\alpha$ — несущественное состояние.*

## *Предельное поведение последовательности $\xi_i$*

Везде в дальнейшем будем предполагать, что случайные величины $\{\eta_n\}_{n\geqslant 1}$ независимы и одинаково распределены на множестве $\mathbb{S}$ с распределением $\mathcal{P} = (p_1, \ldots, p_S)$, причем $p_k > 0$ для всех $k \in \mathbb{S}$.

Обозначим $\tau_k = \tau_k(n) = \sum_{j=1}^{n} \mathbb{I}_k(\xi_j)$, где $\mathbb{I}_A$ — индикатор множества $A$ и $1 \leqslant k \leqslant S$. Кроме того, положим

$$s_j = \sum_{k=1}^{S} p_k \sum_{\substack{\alpha\in\mathfrak{S}_S \\ \alpha_j = k}} \pi_\alpha. \tag{3}$$

Ясно, что $s_k > 0$ и $\sum_k s_k = 1$. Для распределения с вероятностями (3) далее будет использоваться обозначение $\mathcal{R}$.

**Теорема 1** *Для любого начального распределения $\mathcal{L}(\Xi_0)$*

$$\mathrm{P}(\xi_n = k) \underset{n\to+\infty}{\longrightarrow} s_k$$

*и*

$$\tau_k/n \underset{n\to+\infty}{\overset{\mathrm{P}}{\longrightarrow}} s_k, \tag{4}$$

*где $s_k$ — вероятности, определенные в* (3)*.*

## *Эффект выравнивания вероятностей*

Оказывается, что в случае, когда распределение $\mathcal{P} = \mathcal{L}(\eta_i)$ отличается от равномерного $\mathrm{U}_S$, предельное распределение $\mathcal{R}$ последовательности $\{\xi_i\}_{i\geqslant 1}$ оказывается «ближе» к равномерному, чем у исходных $\eta_i$. В качестве меры

близости распределения $\mathcal{Q} = (q_1, \ldots, q_S)$ к равномерному распределению $\mathrm{U}_S$ рассматриваются следующие характеристики:

a) двоичная энтропия

$$\mathcal{H}_2(\mathcal{Q}) = -\sum_{i=1}^{n} q_i \log_2 q_i$$

(чем она больше, тем «ближе» распределение $\mathcal{Q}$ к равномерному),

b) $\rho_1(Q, \mathrm{U}_S) \stackrel{\mathrm{def}}{=} \sum_{k=1}^{S} |q_k - 1/S|$, что представляет собой удвоенное расстояние по вариации между $Q$ и $\mathrm{U}_S$,

c) $\rho_2(Q, \mathrm{U}_S) \stackrel{\mathrm{def}}{=} \sqrt{\sum_{k=1}^{S} (q_k - 1/S)^2}$ и

d) $\rho_\infty(Q, \mathrm{U}_S) \stackrel{\mathrm{def}}{=} \max_k |q_k - 1/S|$.

**Теорема 2** *Имеют место неравенства 1)* $\rho_\infty(\mathcal{R}, \mathrm{U}_S) < \rho_\infty(\mathcal{P}, \mathrm{U}_S)$, *2)* $\rho_q(\mathcal{R}, \mathrm{U}_S) \leqslant \rho_q(\mathcal{P}, \mathrm{U}_S)$ *при* $q \in \{1, 2\}$ *и 3)* $\mathcal{H}_2(\mathcal{R}) > \mathcal{H}_2(\mathcal{P})$.

## *Статистические приложения*

Рассмотрим теперь применение полученных результатов к проверке гипотезы $\mathrm{H}_0$. Для этой цели применим критерий отношения правдоподобия. Зададим при $k \in \mathbb{S}$ величину $\tau_k^{(\xi)} = \sum_{j=1}^{n} \mathbb{I}_k(\xi_j)$, аналогичным образом положим $\tau_k^{(\eta)} = \sum_{j=1}^{n} \mathbb{I}_k(\eta_j)$, и рассмотрим статистики

$$G_n^2(\eta) = 2\sum_{k=1}^{S} \tau_k^{(\eta)} \ln\left(S\tau_k^{(\eta)}/n\right) \quad \text{и} \quad G_n^2(\xi) = 2\sum_{k=1}^{S} \tau_k^{(\xi)} \ln\left(S\tau_k^{(\xi)}/n\right).$$

Хорошо известно, что при выполнении нулевой гипотезы обе статистики асимптотически имеют распределение $\chi^2$ с $S-1$-й степенью свободы. На этом факте и основан критерий отношения правдоподобия, отвергающий гипотезу $\mathrm{H}_0$ при больших значениях $G_n^2(\eta)$ или $G_n^2(\xi)$. В то же время нетрудно видеть, что

$$\widehat{\mathcal{H}}_n(\mathcal{P}) \stackrel{\mathrm{def}}{=} -\sum_{k=1}^{S} \left(\tau_k^{(\eta)}/n\right) \log_2 \left(\tau_k^{(\eta)}/n\right) = \log_2 S - \frac{G_n^2(\eta)}{2n \ln 2}\,,$$

и для выборочной энтропии $\widehat{\mathcal{H}}_n(\mathcal{R})$ выполняется аналогичное тождество. Это означает, что гипотеза $\mathrm{H}_0$ отвергается при слишком маленьких значениях выборочной энтропии.

Поскольку (см. Теорему 2) предельные значения выборочных энтропий $\widehat{\mathcal{H}}_n(\mathcal{P})$ и $\widehat{\mathcal{H}}_n(\mathcal{R})$ удовлетворяют неравенству $\mathcal{H}_2(\mathcal{R}) > \mathcal{H}_2(\mathcal{P})$, то отсюда (и из сходимости (4)) сразу же следует, что при альтернативе $\mathcal{P} \neq \mathrm{U}_S$ критерий отношения правдоподобия, примененный к $\xi_i$, будет при больших $n$ иметь меньшую мощность, чем такой же критерий, примененный к $\eta_i$.

Аналогичные рассуждения можно применить к критерию, основанному на метрике $\rho_\infty$.

С некоторыми оговорками такой же вывод можно сделать относительно критерия $\chi^2$. В этом случае вместо статистик $G_n^2(\eta)$ и $G_n^2(\xi)$ мы имеем дело с

$$\chi_n^2(\eta) = \sum_{k=1}^{S} \frac{\left(\tau_k^{(\eta)} - n/S\right)^2}{n/S} \quad \text{и} \quad \chi_n^2(\xi) = \sum_{k=1}^{S} \frac{\left(\tau_k^{(\xi)} - n/S\right)^2}{n/S},$$

причем нулевая гипотеза отвергается, если статистика $\chi_n^2$ оказывается слишком большой. Заметим, что

$$\chi_n^2(\eta)/(Sn) = \sum_{k=1}^{S} \left(\tau_k^{(\eta)}/n - 1/S\right)^2 \xrightarrow{\mathrm{P}} \rho_2(\mathcal{P}, \mathrm{U}_S),$$

для статистики $\chi_n^2(\xi)/(Sn)$ имеет место аналогичная сходимость к $\rho_2(\mathcal{R}, \mathrm{U}_S)$, причем $\rho_2(\mathcal{P}, \mathrm{U}_S) \geqslant \rho_2(\mathcal{R}, \mathrm{U}_S)$. Поэтому можно ожидать, что при больших $n$ с вероятностью, близкой к 1, значение статистики $\chi_n^2(\eta)$ будет больше, чем значение статистики $\chi_n^2(\xi)$ — по крайней мере для тех распределений $\mathcal{P}$, для которых $\rho_2(\mathcal{P}, \mathrm{U}_S) > \rho_2(\mathcal{R}, \mathrm{U}_S)$.

Подобные рассуждения годятся и для критерия, основанного на метрике $\rho_1$.

## Литература


[1] Рябко Б.Я., Сжатие данных с помощью стопки книг // Проблемы передачи информации, 1980, Т. XVI, Вып. 4, С. 16–20.

[2] A locally adaptive data compression scheme / Jon Louis Bentley, Daniel D. Sleator, Robert E. Tarjan, Victor K. Wei // Commun. ACM., 1986., Vol. 29, no. 4.

[3] Рябко Б.Я., Пестунов А.И., “Стопка книг” как новый статистический тест для случайных чисел, // Пробл. передачи информ., 2004, Т. 40,

Вып. 1, С. 73–78.

[4] Read T., Cressie N., Goodness-of-Fit Statistics for Discrete Multivariate Data, Springer-Verlag, New York, 1988.