Math-Net.Ru

Общероссийский математический портал

А. Л. Рухин, Совместное распределение частот цепочек и многомерный закон
Пойа–Эпли, *Теория вероятн. и ее примен.*, 2009, том 54, выпуск 2, 226–242

DOI: <https://doi.org/10.4213/tvp2697>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и
согласны с пользовательским соглашением

<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 54.224.135.184

26 апреля 2023 г., 12:45:36

© 2009 г.

РУХИН А. Л.*

СОВМЕСТНОЕ РАСПРЕДЕЛЕНИЕ ЧАСТОТ ЦЕПОЧЕК И МНОГОМЕРНЫЙ ЗАКОН ПОЙА–ЭПЛИ¹⁾

В статье изучается совместное распределение частот перекрывающихся слов в марковской последовательности. Обычно характеристики этого распределения выражаются в терминах так называемой корреляционной матрицы цепочек. Более непосредственный подход дает явные формулы, использующие фундаментальную матрицу марковской цепи, состояниями которой являются слова заданной длины. Эти формулы приводят к производящей функции вероятностей асимптотического распределения частот, соответствующей новому многомерному дискретному закону.

Ключевые слова и фразы: корреляционная матрица цепочек, полиномы Белла, сложное распределение Пуассона, соседства, формула Лагранжа, фундаментальная матрица, цепь Маркова.

1. Введение. Рассматривается случайный текст, образованный последовательностью букв из конечного алфавита. Для определенного набора слов (или цепочек) представляет интерес вероятность заданного числа их появлений в этом тексте. Такая задача возникает в теории информации (входное кодирование, проверка случайности) и в молекулярной биологии (анализ ДНК, распознавание генов). В [18, гл. 7] обсуждаются теоретико-информационные аспекты. Элементарное введение в анализ частот в последовательности ДНК дано в [13]. В предположении независимости соответствующие вероятности выражаются через так называемый корреляционный полином, введенный в [9].

В п. 2 для регулярной цепи Маркова обсуждается корреляционная матрица цепочек и указана ее связь с фундаментальной матрицей марковской цепи, состояниями которой являются слова заданной длины. Указан вид производящей функции вероятностей частот слов. В п. 3 получено асимптотическое распределение этих частот в виде нового многомерного сложного закона Пуассона. Это распределение, названное многомерным законом Пойа–Эпли, обсуждается в п. 4. В заключительном п. 5 приводятся несколько примеров.

---

*National Institute of Standards and Technology, SED, Gaithersburg MD, 20899, USA;
e-mail: andrew.rukhin@nist.gov

¹⁾ Работа выполнена при поддержке гранта № H98230-06-1-0068.

Упомянутое асимптотическое распределение получено в предположении, что длина цепочек m , так же как и порядок цепи Маркова, неограниченно возрастают. Схожее условие использовалось в [19], [20], [8] и [3].

2. Корреляционные матрицы цепочек. Рассмотрим случайный текст $X_n = (\varepsilon_1, \dots, \varepsilon_n)$, образованный стационарной последовательностью дискретных случайных величин ε_k , принимающих значения в алфавите $\mathbb{Q} = \{1, \dots, q\}$ при некотором $q \geq 2$. Начнем с марковской цепи порядка m , где $m \geq 1$. Многие статистические процедуры анализа такой последовательности основаны на частотах перекрывающихся слов (цепочек) длины m типа $i^m = (i_1 \dots i_m)$.

Множество всех таких слов мощности q^m обозначается $\mathbb{Q}^m$. Предполагается, что для любого $i = i^m \in \mathbb{Q}^m$ вероятность $P(i) = \mathbf{P}((\varepsilon_1, \dots, \varepsilon_m) = (i_1 \dots i_m))$ положительна. Пусть $P(j_{k+1} \dots j_m | i) = P(ij_{k+1} \dots j_m) / P(i)$ означает условную вероятность того, что слову $j_{k+1} \dots j_m$, $k = 1, \dots, m$, предшествует слово i . Обозначим через $\mathbb{I}$ единичную матрицу, через e — вектор размерности q^m с единичными координатами и через p q^m -мерный вектор с координатами $P(i)$, $i \in \mathbb{Q}^m$. Матрица $\mathbb{P} = \mathbb{P}_m$ размера $q^m \times q^m$ образована элементами

$$P_{ij} = \delta_{(i_2 \dots i_m), (j_1 \dots j_{m-1})} \mathbf{P}(\varepsilon_{n+m} = j_m | \varepsilon_n = i_1, \dots, \varepsilon_{n+m-1} = i_m). \quad (1)$$

Здесь и далее $\delta_{\ell, \kappa} = \prod_{k=1}^K \delta_{l_k \kappa_k}$ — символ Кронекера для двух K -индексов $\ell = (\ell_1, \dots, \ell_K)$ и $\kappa = (\kappa_1, \dots, \kappa_K)$.

Корреляционная матрица цепочек $\mathbb{C}(z)$ определена для комплексных z , для которых существует $[\mathbb{I} - z(\mathbb{P} - ep^T)]^{-1}$:

$$\begin{aligned} \mathbb{C}(z) &= [\mathbb{I} - z(\mathbb{P} - ep^T)]^{-1} + (z + \dots + z^{m-1})ep^T \\ &= \left[\mathbb{I} - z\mathbb{P} + \frac{z^m}{1 + \dots + z^{m-1}} ep^T \right]^{-1}. \end{aligned} \quad (2)$$

Эта матрица использовалась в [18] для последовательностей Бернулли. Она приводит к производящей функции вероятностей частот слов в случайном тексте [16], [17].

Корреляционный полином цепочек $C_{ij}(z)$ был определен в [9] для двух слов $i = (i_1 \dots i_m)$ и $j = (j_1 \dots j_m)$ следующим образом:

$$C_{ij}(z) = \sum_{r=0}^{m-1} \delta_{(i_{r+1} \dots i_m), (j_1 \dots j_{m-r})} P(j_{m-r+1} \dots j_m | i) z^r. \quad (3)$$

Если $C(z)$ обозначает матрицу, образованную корреляционными полиномами цепочек $C_{ij}(z)$, то при $0 \leq z < 1$

$$\mathbb{C}(z) = C(z) + \sum_{k=m}^{\infty} z^k [\mathbb{P} - ep^T]^k = [\mathbb{I} - z\mathbb{P}]^{-1} - \frac{z^m}{1 - z} ep^T,$$

поскольку $\mathbb{P}^T p = p$ и $\mathbb{P} e = e$. В случае независимых одинаково распределенных величин выполнено равенство $\mathbb{C}(z) = C(z)$. Тогда все собственные значения матрицы $\mathbb{P}$ равны 1 (с кратностью один) или 0 (с кратностью $q^m - 1$), так что $\mathbb{C}(z)$ определена при всех z . Согласно определению,

$$\mathbb{C}(1) = \left[\mathbb{I} - \mathbb{P} + \frac{1}{m} e p^T \right]^{-1}.$$

Хотя корреляционные полиномы цепочек $C_{ij}(z)$ адекватны задаче изучения независимых одинаково распределенных величин, в общем случае марковской зависимости нужна корреляционная матрица цепочек. Матрица $\mathbb{C}(1)$ играет важнейшую роль, когда Y_n есть цепь Маркова порядка один, состояния которой — слова длины m , т.е. $Y_n = i$, если $(\varepsilon_n, \dots, \varepsilon_{n+m-1}) = (i_1 \dots i_m)$. Для этой цепи переходные вероятности имеют вид (1). Мы предполагаем, что матрица переходных вероятностей $\mathbb{P}$ имеет единственное собственное число с единичным модулем, т.е. цепь регулярная или эргодическая. Это условие выполнено, если все вероятности $P(i^{m+1})$ положительны, и тогда $\rho(\mathbb{P} - e p^T) < 1$, где $\rho(A)$ обозначает спектральный радиус матрицы A . Частота $\omega_i = \omega_i(n)$ появления слова i в начальной последовательности совпадает с числом посещений состояния i цепью Y_n .

Хорошо известно, что асимптотическое совместное распределение частот нормально. Для нашей матрицы $\mathbb{P}$ стационарное распределение задается вектором p с координатами $P(i) = P(\varepsilon_1 = i_1, \dots, \varepsilon_m = i_m)$. Фундаментальная матрица $\mathbb{Z}$ цепи Y_n определяется формулой $\mathbb{Z} = (\mathbb{I} - \mathbb{P} + e p^T)^{-1}$. Эта матрица допускает простое выражение через $\mathbb{C}(1)$:

$$\mathbb{Z} = \mathbb{C}(1) - (m - 1) e p^T. \quad (4)$$

Важным свойством фундаментальной матрицы является то, что она задает ковариационную матрицу предельного распределения частот. Согласно теореме 4.6.1 в [10], для фиксированного m предельная ковариационная матрица $\Sigma = \lim_{n \rightarrow \infty} n^{-1} \text{cov}(\omega_i, \omega_j)$ имеет элементы

$$\sigma_{ij} = -\delta_{ij} P(i) - P(i) P(j) + P(i) \mathbb{Z}_{ij} + P(j) \mathbb{Z}_{ji}. \quad (5)$$

В силу (4) матрица Σ может быть записана в виде

$$\Sigma = -\mathbb{D} - (2m - 1) p p^T + \mathbb{D} \mathbb{C}(1) + \mathbb{C}(1)^T \mathbb{D}$$

с диагональной матрицей $\mathbb{D} = \text{diag}(p)$.

Таким образом, предельная ковариационная матрица совместного распределения частот m -слов в марковской последовательности порядка m просто выражается через корреляционную матрицу цепочек $\mathbb{C}$.

3. Производящие функции вероятностей частот и их асимптотика. Начнем со следующего результата, дающего удобное представление производящей функции частот $\omega_i(n)$, $i \in \mathbb{Q}^m$.

Лемма 1. Пусть $X_n = (\varepsilon_1, \dots, \varepsilon_n)$ — цепь Маркова порядка m , а $\omega_i(n)$, $i \in \mathbb{Q}^m$, — частоты перекрывающихся m -слов, $\sum_i \omega_i = n - m + 1 = N$. Производящая функция совместного распределения этих частот имеет вид

$$\begin{aligned} \mathbf{E} \prod_i u_i^{\omega_i} &= 1 + N \sum_{\ell} (u_{\ell} - 1) P(\ell) \\ &+ \sum_{s=1}^{N-1} \sum_{\ell_0, \dots, \ell_s} \prod_{k=0}^s (u_{\ell_k} - 1) P(\ell_0) (e_{\ell_0} \otimes \dots \otimes e_{\ell_{s-1}})^T \\ &\quad \times \mathbb{T}_N(s) (e_{\ell_1} \otimes \dots \otimes e_{\ell_s}), \end{aligned} \quad (6)$$

где $\ell_0, \dots, \ell_s$ пробегают множество $\mathbb{Q}^{m(s+1)}$, e_{ℓ} обозначает ℓ -й базисный вектор, $\ell \in \mathbb{Q}^m$, а

$$\mathbb{T}_N(s) = \sum_{\substack{i_1, \dots, i_s \geq 1 \\ i_1 + \dots + i_s \leq N}} (N - i_1 - \dots - i_s) \mathbb{P}^{i_1} \otimes \dots \otimes \mathbb{P}^{i_s}. \quad (7)$$

Здесь $\mathbb{P}_1 \otimes \mathbb{P}_2$ обозначает тензорное произведение матриц (или векторов) $\mathbb{P}_1$ и $\mathbb{P}_2$, а матрица $\mathbb{P} = \mathbb{P}_m$ определена в (1).

Доказательство. Чтобы вывести это представление, применим известную формулу для производящей функции частот марковской цепи [6, с. 41], согласно которой для диагональной матрицы $\mathbb{U}$ с диагональными элементами u_i , $i \in \mathbb{Q}^m$,

$$M_N(\mathbb{U}) = \mathbf{E} \prod_i u_i^{\omega_i} = p^T \mathbb{U} (\mathbb{P} \mathbb{U})^{N-1} e. \quad (8)$$

Для матрицы $\tilde{\mathbb{P}} = \mathbb{P}(\mathbb{U} - \mathbb{I})$ имеем

$$(\mathbb{P} \mathbb{U})^{N-1} = (\mathbb{P} + \tilde{\mathbb{P}})^{N-1} = \sum_{s=0}^{N-1} \sum_{\substack{i_1, \dots, i_{s+1} \geq 0 \\ i_1 + \dots + i_{s+1} = N-1-s}} \mathbb{P}^{i_1} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_{s+1}},$$

так что в члены со степенями $i_1, \dots, i_{s+1}$ матрица $\tilde{\mathbb{P}}$ входит s раз. Тогда $\mathbb{P}^{i_{s+1}} e = e$, $\tilde{\mathbb{P}} e = \sum_{\ell} (u_{\ell} - 1) \mathbb{P} e_{\ell}$ и

$$\begin{aligned} (\mathbb{P} \mathbb{U})^{N-1} e &= \sum_{s=0}^{N-1} \sum_{\substack{i_1, \dots, i_s \geq 0 \\ i_1 + \dots + i_s \leq N-1-s}} \mathbb{P}^{i_1} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_s} \tilde{\mathbb{P}} e \\ &= e + \sum_{\ell} (u_{\ell} - 1) \sum_{s=1}^{N-1} \sum_{\substack{i_1, \dots, i_s \geq 0 \\ i_1 + \dots + i_s \leq N-1-s}} \mathbb{P}^{i_1} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_s+1} e_{\ell}. \end{aligned}$$

Поскольку $p^T \mathbb{P}^{i_1} \tilde{\mathbb{P}} = p^T \tilde{\mathbb{P}} = p^T (\mathbb{U} - \mathbb{I})$, имеем

$$\begin{aligned} p^T (\mathbb{P}\mathbb{U})^{N-1} e &= 1 + (N-1) \sum_{\ell} (u_{\ell} - 1) p^T e_{\ell} \\ &+ \sum_{\ell} (u_{\ell} - 1) \sum_{s=2}^{N-1} \sum_{\substack{i_2, \dots, i_s \geq 0 \\ i_2 + \dots + i_s \leq N-1-s}} (N-s-i_2-\dots-i_s) p^T \tilde{\mathbb{P}} \mathbb{P}^{i_2} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_s+1} e_{\ell} \\ &= 1 + (N-1) \sum_{\ell} (u_{\ell} - 1) P(\ell) + \sum_{j, \ell} (u_j - 1)(u_{\ell} - 1) P(j) \\ &\times \sum_{s=1}^{N-2} \sum_{\substack{i_1, \dots, i_s \geq 0 \\ i_1 + \dots + i_s \leq N-2-s}} (N-1-s-i_1-\dots-i_s) e_j^T \mathbb{P}^{i_1} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_s+1} e_{\ell}. \end{aligned}$$

Аналогично,

$$\begin{aligned} p^T (\mathbb{U} - \mathbb{I})(\mathbb{P}\mathbb{U})^{N-1} e &= \sum_{\ell} (u_{\ell} - 1) P(\ell) \\ &+ \sum_{j, \ell} (u_j - 1)(u_{\ell} - 1) P(j) \sum_{s=1}^{N-1} \sum_{\substack{i_1, \dots, i_s \geq 0 \\ i_1 + \dots + i_s \leq N-1-s}} e_j^T \mathbb{P}^{i_1} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_s+1} e_{\ell}. \end{aligned}$$

Таким образом,

$$\begin{aligned} p^T \mathbb{U}(\mathbb{P}\mathbb{U})^{N-1} e &= 1 + N \sum_{\ell} (u_{\ell} - 1) P(\ell) + \sum_{j, \ell} (u_{\ell} - 1)(u_j - 1) P(j) \\ &\times \sum_{s=1}^{N-1} \sum_{\substack{i_1, \dots, i_s \geq 0 \\ i_1 + \dots + i_s \leq N-1-s}} (N-s-i_1-\dots-i_s) e_j^T \mathbb{P}^{i_1} \tilde{\mathbb{P}} \dots \tilde{\mathbb{P}} \mathbb{P}^{i_s+1} e_{\ell}. \end{aligned}$$

При $k = 1, 2, \dots, s-1$ мы заменим k -й множитель $\tilde{\mathbb{P}}$ в этой формуле на $\sum_{\ell_k} (u_{\ell_k} - 1) \mathbb{P} e_{\ell_k} e_{\ell_k}^T$, положим $\ell_0 = j$, $\ell_s = \ell$ и подставим в (8). Получаем

$$\begin{aligned} M_N(\mathbb{U}) &= 1 + N \sum_{\ell} (u_{\ell} - 1) P(\ell) \\ &+ \sum_{s=1}^{N-1} \sum_{\ell_0, \dots, \ell_s} \prod_{k=0}^s (u_{\ell_k} - 1) P(\ell_0) (e_{\ell_0} \otimes \dots \otimes e_{\ell_{s-1}})^T \mathbb{T}_N(s) (e_{\ell_1} \otimes \dots \otimes e_{\ell_s}). \end{aligned}$$

Лемма 1 доказана.

Мы применим лемму 1 в ситуации, когда слова принадлежат заданному множеству $\Omega_m \subset \mathbb{Q}^m$. Тогда сумма в (6) распространяется на $\ell_0, \dots, \ell_s \in \Omega_m$.

Наш главный результат относится к последовательности m -слов $\omega_{im} = \omega_{im}(n)$, $i^m \in \mathbb{Q}^m$. В совместном асимптотическом распределении частот $n \rightarrow \infty$, $m = m(n) \rightarrow \infty$, $N/n \rightarrow 1$. Предполагается, что Ω — конечное множество, образованное счетными последовательностями

$\iota = (\iota_1, \dots, \iota_m, \dots)$. Множество Ω_m , $m = 1, 2, \dots$, есть сужение Ω , т.е. состоит из m -цепочек $\iota^m = (\iota_1, \dots, \iota_m)$.

Для простоты обозначений пусть теперь $\mathbb{P}$ обозначает подматрицу с элементами (1), отвечающими строкам и столбцам в Ω_m ; $\mathbb{I}$ обозначает единичную матрицу соответствующего размера, и то же соглашение принимаем для векторов e и p .

Теорема 1. Пусть $\{X_n = (\varepsilon_1, \dots, \varepsilon_n), n = 1, 2, \dots\}$ — последовательность эргодических цепей Маркова, где X_n имеет порядок $m = m(n)$. Предположим, что $n \rightarrow \infty$ так, что $m \rightarrow \infty$, $m/n \rightarrow 0$, и для всех $\iota \in \Omega$

$$\lim_{n \rightarrow \infty} nP(\iota^m) = \frac{\lambda}{\beta_\iota}, \quad (9)$$

где λ и β_ι положительны. Предположим также, что

$$\sup_m \rho(\mathbb{P} - ep^T) < 1, \quad (10)$$

$$\mathbb{C}_{\Omega_m}(1) \rightarrow (\mathbb{I} - \Pi)^{-1} \quad (11)$$

с некоторой субстохастической матрицей Π такой, что для диагональной матрицы $\mathbb{B}$, образованной вектором β_ι , матрица $\mathbb{B}\Pi^T\mathbb{B}^{-1}$ тоже является субстохастической и

$$e^T\mathbb{B}^{-1}(\mathbb{I} - \Pi)e = 1. \quad (12)$$

Тогда совместное распределение частот ω_{ι^m} , $\iota^m \in \Omega_m$, слабо сходится к многомерному сложному закону Пуассона (закону Пойа–Энли):

$$\mathbf{E} \prod_{\iota \in \Omega} u_\iota^{\omega_{\iota^m}} \rightarrow \exp\{\lambda[e^T\mathbb{B}^{-1}(\mathbb{I} - \Pi)(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e - 1]\}. \quad (13)$$

Д о к а з а т е л ь с т в о. Чтобы применить лемму 1, отметим, что для фиксированных m , $s = 1, \dots, N - 1$ и $\ell_0, \dots, \ell_s \in \Omega$

$$e_{\ell_k^m}^T \mathbb{P}^{i_{k+1}} e_{\ell_{k+1}^m} = e_{\ell_k^m}^T (\mathbb{P} - ep^T)^{i_{k+1}} e_{\ell_{k+1}^m} + P(\ell_{k+1}^m)$$

при любых положительных $i_1, \dots, i_s$ таких, что $i_1 + \dots + i_s \leq N - 1$. Таким образом,

$$\begin{aligned} & (e_{\ell_0^m} \otimes \dots \otimes e_{\ell_{s-1}^m})^T \mathbb{P}^{i_1} \otimes \dots \otimes \mathbb{P}^{i_s} (e_{\ell_1^m} \otimes \dots \otimes e_{\ell_s^m}) \\ &= \sum_{r=0}^s \sum_{\substack{t_1, \dots, t_r \\ t_{r+1}, \dots, t_s}}^r \prod_{k=1}^r e_{\ell_{t_k}^m}^T (\mathbb{P} - ep^T)^{i_{t_k+1}} e_{\ell_{t_k+1}^m} P(\ell_{t_{r+1}+1}^m) \dots P(\ell_{t_s+1}^m). \end{aligned}$$

Здесь $t_1, \dots, t_s$, $t_1 < \dots < t_r$, — перестановка чисел $0, 1, \dots, s - 1$.

В силу (11), для всех j и ℓ из Ω

$$\sum_{i \geq 1} e_{\ell^m}^T (\mathbb{P} - ep^T)^i e_{j^m} = e_{\ell^m}^T (\mathbb{I} - \mathbb{P} + ep^T)^{-1} e_{j^m} - \delta_{\ell^m j^m} \rightarrow \frac{\alpha_{\ell j}}{\beta_j},$$

где α_{ℓ_j} обозначают элементы матрицы $\mathbb{A} = \Pi[\mathbb{I} - \Pi]^{-1}\mathbb{B} = [\mathbb{I} - \Pi]^{-1}\mathbb{B} - \mathbb{B}$. Согласно (10), при $i \rightarrow \infty$ последовательность $e_{\ell^m}^T(\mathbb{P} - ep^T)^i e_{j^m}$ сходится к нулю равномерно по m , так что этот ряд тоже сходится равномерно по m .

Для заданных s и $1 \leq r \leq s$ находим, что

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{i_{t_1+1}, \dots, i_{t_r+1} \geq 1} \prod_{k=1}^r e_{\ell_{t_k}^m}^T (\mathbb{P} - ep^T)^{i_{t_k}+1} e_{\ell_{t_k}^m+1} \\ &= \lim_{n \rightarrow \infty} \prod_{k=1}^r \left(e_{\ell_{t_k}^m}^T (\mathbb{I} - \mathbb{P} + ep^T)^{-1} e_{\ell_{t_k}^m+1} - \delta_{\ell_{t_k}^m, \ell_{t_k}^m+1} \right) = \frac{\alpha_{\ell_1 \ell_2} \cdots \alpha_{\ell_r \ell_{r+1}}}{\beta_{\ell_2} \cdots \beta_{\ell_{r+1}}}. \end{aligned}$$

Поэтому (с $\mathbb{T}_N(s)$, определенным в (7)) получаем, что

$$\begin{aligned} & N^{-1} (e_{\ell_0^m} \otimes \cdots \otimes e_{\ell_{s-1}^m})^T \mathbb{T}_N(s) (e_{\ell_1^m} \otimes \cdots \otimes e_{\ell_s^m}) \\ &= N^{-1} \sum_{r=0}^s \sum_{t_1, \dots, t_s} \sum_{\substack{i_{t_1+1}, \dots, i_{t_s+1} \geq 1 \\ i_{t_1+1} + \dots + i_{t_s+1} \leq N-1}} (N - i_{t_1+1} - \cdots - i_{t_s+1}) \\ &\quad \times \prod_{k=1}^r e_{\ell_{t_k}^m}^T (\mathbb{P} - ep^T)^{i_{t_k}+1} e_{\ell_{t_k}^m+1} P(\ell_{t_{r+1}+1}^m) \cdots P(\ell_{t_s+1}^m) \\ &\sim \frac{\lambda^{s-r}}{N^{s-r+1}} \sum_{r=0}^s \sum_{\substack{t_1 < \dots < t_r \\ t_{r+1} < \dots < t_s}} \prod_{k=r+1}^s \beta_{\ell_{t_k}+1}^{-1} \\ &\quad \times \sum_{\substack{i_{t_1+1}, \dots, i_{t_r+1} \geq 1 \\ i_{t_1+1} + \dots + i_{t_r+1} \leq N-s+r-1}} \binom{N - i_{t_1+1} - \cdots - i_{t_r+1}}{s-r+1} \\ &\quad \times \prod_{k=1}^r e_{\ell_{t_k}^m}^T (\mathbb{P} - ep^T)^{i_{t_k}+1} e_{\ell_{t_k}^m+1} \\ &\rightarrow \sum_{r=0}^s \frac{\lambda^{s-r}}{(s-r+1)! \prod_{k=1}^s \beta_{\ell_k}} \sum_{t_1 < \dots < t_r} \alpha_{\ell_{t_1} \ell_{t_1+1}} \cdots \alpha_{\ell_{t_r} \ell_{t_r+1}} \\ &= \sum_{r=0}^s \frac{\lambda^{s-r}}{(s-r+1)! \prod_{k=1}^s \beta_{\ell_k}} \mathbb{S}_r(\alpha_{\ell_0 \ell_1}, \dots, \alpha_{\ell_{s-1} \ell_s}), \end{aligned}$$

где $\mathbb{S}_r$ обозначает r -ю элементарную симметрическую функцию:

$$\mathbb{S}_r(a_1, \dots, a_s) = \sum_{1 \leq i_1 < \dots < i_r \leq s} a_{i_1} \cdots a_{i_r}.$$

Действительно, если $i_{u_1}, \dots, i_{u_{s-r}}$ обозначают индексы, отличные от $i_{t_1+1}, \dots, i_{t_r+1}$, то, полагая $I = \sum (i_{t_1+1} + \dots + i_{t_r+1})$, имеем

$$\sum_{\substack{i_{u_1}, \dots, i_{u_{s-r}} \geq 1 \\ i_{u_1} + \dots + i_{u_{s-r}} \leq N-I-1}} (N - I - i_{u_1} - \cdots - i_{u_{s-r}}) = \binom{N-I}{s-r+1},$$

и в силу того, что ряд $\sum_{i_{t_1+1}, \dots, i_{t_r+1} \geq 1} \prod_{k=1}^r e_{\ell_{t_k}}^T (\mathbb{P} - ep^T)^{i_{t_k+1}} e_{\ell_{t_k+1}}^m$ сходится равномерно, получаем, что

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{(s-r+1)!}{N^{s-r+1}} \sum_{\substack{i_{t_1+1}, \dots, i_{t_r+1} \geq 1 \\ i_{t_1+1} + \dots + i_{t_r+1} \leq N-s+r-1}} \binom{N-i_{t_1+1}-\dots-i_{t_r+1}}{s-r+1} \\ & \times \prod_{k=1}^r e_{\ell_{t_k}}^T (\mathbb{P} - ep^T)^{i_{t_k+1}} e_{\ell_{t_k+1}}^m \\ & = \lim_{n \rightarrow \infty} \left[\binom{N}{s-r+1} \right]^{-1} \sum_{i=0}^{N-s} \binom{N-i-r}{s-r} \\ & \times \sum_{i_{t_1} + \dots + i_{t_r} \leq i+r-1} \prod_{k=1}^r e_{\ell_{t_k}}^T (\mathbb{P} - ep^T)^{i_{t_k+1}} e_{\ell_{t_k+1}}^m = \frac{\alpha_{\ell_{t_1} \ell_{t_1+1}} \cdots \alpha_{\ell_{t_r} \ell_{t_r+1}}}{\beta_{\ell_{t_1+1}} \cdots \beta_{\ell_{t_r+1}}}. \end{aligned}$$

Из (6) следует, что $\lim_{n \rightarrow \infty} (M_N(\mathbb{U})/M(\mathbb{U})) = 1$, где для $v_\ell = \beta_\ell^{-1}(u_\ell - 1)$

$$\begin{aligned} M(\mathbb{U}) &= 1 + \sum_{s=0}^{\infty} \sum_{\ell_0, \dots, \ell_s} \prod_{k=0}^s v_{\ell_k} \sum_{r=0}^s \frac{\lambda^{s-r+1}}{(s-r+1)!} \mathbb{S}_r(\alpha_{\ell_0 \ell_1}, \dots, \alpha_{\ell_{s-1} \ell_s}) \\ &= 1 + \sum_{r=0}^{\infty} \sum_{s=r}^{\infty} \frac{\lambda^{s-r+1}}{(s-r+1)!} \sum_{\ell_0, \dots, \ell_s} \prod_{k=0}^s v_{\ell_k} \mathbb{S}_r(\alpha_{\ell_0 \ell_1}, \dots, \alpha_{\ell_{s-1} \ell_s}) \\ &= 1 + \sum_{r=0}^{\infty} \sum_{s=r}^{\infty} \frac{\lambda^{s-r+1}}{(s-r+1)!} \sum_{\ell_0, \dots, \ell_s} \prod_{k=0}^s v_{\ell_k} \sum_{t_1 < \dots < t_r} \alpha_{\ell_{t_1} \ell_{t_1+1}} \cdots \alpha_{\ell_{t_r} \ell_{t_r+1}}. \quad (14) \end{aligned}$$

В сумме (14) $\ell_0, \dots, \ell_s \in \Omega$; для ее вычисления разобьем множество $\{t_1 < \dots < t_r\}$ согласно числу соседств заданной длины. Более точно, назовем последовательность $t_{i_1} < \dots < t_{i_p}$ *соседством* длины p , если $t_{i_{k+1}} = t_{i_k} + 1$ для $k = 1, \dots, p-1$, но $t_{i_{p+1}} \neq t_{i_p} + 1$ и $t_{i_1} \neq t_{i_1-1} + 1$. (Если $t_1 = 0$ или $t_p = r$, будем считать эти неравенства выполненными автоматически.) Перечисление последовательностей с заданными числами соседств обсуждается в [7, п. 2.3.15].

Сумма по индексам, образующим соседство длины p , легко вычисляется:

$$\begin{aligned} & \sum_{t_{i_1} < \dots < t_{i_p}} \alpha_{\ell_{t_{i_1}} \ell_{t_{i_1}+1}} \alpha_{\ell_{t_{i_2}} \ell_{t_{i_2}+1}} \cdots \alpha_{\ell_{t_{i_p}} \ell_{t_{i_p}+1}} v_{\ell_{t_{i_1}}} v_{\ell_{t_{i_2}}} \cdots v_{\ell_{t_{i_p}}} v_{\ell_{t_{i_p}+1}} \\ & = v^T \mathbb{A}(\mathbb{V} \mathbb{A})^{p-1} v =: z_p. \end{aligned} \quad (15)$$

Здесь $\mathbb{V}$ обозначает диагональную матрицу, образованную вектором v с координатами $v_\ell = \beta_\ell^{-1}(u_\ell - 1)$.

При фиксированном s обозначим $C_s(\nu_1, \dots, \nu_r)$ число последовательностей $t_1 < \dots < t_r$, образованных символами $0, \dots, s-1$ с ν_p соседствами

длины p , $p = 1, \dots, r$. Ясно, что $\sum p\nu_p = r$ и

$$\begin{aligned} & \sum_{\ell_0, \dots, \ell_s} \prod_{k=0}^s v_{\ell_k} \sum_{t_1 < \dots < t_r} \alpha_{\ell_{t_1} \ell_{t_1+1}} \dots \alpha_{\ell_{t_r} \ell_{t_r+1}} \\ &= \sum_{\substack{\nu_1, \dots, \nu_r \\ \sum p\nu_p = r}} (v^T e)^{s - \sum \nu_p - r + 1} C_s(\nu_1, \dots, \nu_r) z_1^{\nu_1} \dots z_r^{\nu_r}. \end{aligned}$$

Мы докажем, что

$$C_s(\nu_1, \dots, \nu_r) = \binom{s - r + 1}{\sum \nu_p} \binom{\sum \nu_p}{\nu_1 \dots \nu_r}. \quad (16)$$

В самом деле, если $\sum \nu_p = h$, то чтобы построить r -множество $t_1 < \dots < t_r$ из чисел $\{0, 1, \dots, s\}$, выберем сначала $h + 1$ положительных чисел $a_0, a_1, \dots, a_h$ с суммой $s - r + 2$. Это можно сделать $\binom{s - r + 1}{h}$ способами. Уменьшим a_0 и a_h на единицу, так что новая сумма равна $s - r$, а a_0 и a_m — возможно, нули. Затем расположим ν_1 единиц, ν_2 двоек, $\dots$, ν_r символов r в виде последовательности $(b_1, \dots, b_r)$. Для этого имеется $\binom{h}{\nu_1 \dots \nu_r}$ способов. Теперь получим множество из r элементов следующим образом: пропускаем $1, \dots, a_0$; последующие b_1 символов включаются в наше множество, следующие a_1 символов пропускаются, а следующие b_2 символов включаются и т.д. Формула (16) перечисляет все такие множества.

Для z_p , определенного в (15), получаем

$$\begin{aligned} M(\mathbb{U}) &= 1 + \sum_{r=0}^{\infty} \sum_{s=r}^{\infty} \frac{\lambda^{s-r+1} (v^T e)^{s-r+1}}{(s - r + 1)!} \\ &\quad \times \sum_{\substack{\nu_1, \dots, \nu_r \\ \sum p\nu_p = r}} (v^T e)^{-\sum \nu_p} z_1^{\nu_1} \dots z_r^{\nu_r} C_s(\nu_1, \dots, \nu_r) \\ &= e^{\lambda v^T e} + \sum_{r=1}^{\infty} \sum_{\substack{\nu_1, \dots, \nu_r \\ \sum p\nu_p = r}} \frac{\lambda^{\sum \nu_p} z_1^{\nu_1} \dots z_r^{\nu_r}}{\nu_1! \dots \nu_r!} \sum_{s=r+\sum \nu_p-1}^{\infty} \frac{(\lambda v^T e)^{s-\sum \nu_p+r+1}}{(s - \sum \nu_p + r + 1)!} \\ &= e^{\lambda v^T e} \sum_{r=0}^{\infty} \sum_{\substack{\nu_1, \dots, \nu_r \\ \sum p\nu_p = r}} \frac{\lambda^{\sum \nu_p} z_1^{\nu_1} \dots z_r^{\nu_r}}{\nu_1! \dots \nu_r!}. \end{aligned}$$

Сумма

$$\sum_{\substack{\nu_1, \dots, \nu_r \\ \sum p\nu_p = h}} \frac{\lambda^h r! z_1^{\nu_1} \dots z_r^{\nu_r}}{\nu_1! \dots \nu_r!}$$

задает *полином Белла* $Y_r(\lambda; z_1, 2!z_2, \dots, r!z_r)$. Поскольку $\sum_{k=1}^{\infty} z_k t^k = tv^T \mathbb{A}(\mathbb{I} - t\mathbf{V}\mathbb{A})^{-1}v$, формула (45) в [12] показывает, что

$$\sum_{r=0}^{\infty} \sum_{\substack{\nu_1, \dots, \nu_r \\ \sum \nu_p = r}} \frac{\lambda \sum \nu_p z_1^{\nu_1} \dots z_r^{\nu_r}}{\nu_1! \dots \nu_r!} = \exp \{ \lambda v^T \mathbb{A}(\mathbb{I} - \mathbf{V}\mathbb{A})^{-1}v \}.$$

Так как $\mathbf{V} = (\mathbb{U} - \mathbb{I})\mathbb{B}^{-1}$, получаем

$$\begin{aligned} \mathbb{I} - \mathbf{V}\mathbb{A} &= \mathbb{I} + \mathbb{B}^{-1}\Pi(\mathbb{I} - \Pi)^{-1}\mathbb{B} - \mathbb{B}^{-1}\mathbb{U}\Pi(\mathbb{I} - \Pi)^{-1}\mathbb{B} \\ &= \mathbb{B}^{-1}(\mathbb{I} - \Pi)^{-1}\mathbb{B} - \mathbb{B}^{-1}\mathbb{U}\Pi(\mathbb{I} - \Pi)^{-1}\mathbb{B} = \mathbb{B}^{-1}(\mathbb{I} - \mathbb{U}\Pi)(\mathbb{I} - \Pi)^{-1}\mathbb{B}. \end{aligned}$$

Поскольку $v = (\mathbb{U} - \mathbb{I})\mathbb{B}^{-1}e = \mathbb{B}^{-1}(\mathbb{U} - \mathbb{I})e$, имеем

$$\begin{aligned} v^T \mathbb{A}(\mathbb{I} - \mathbf{V}\mathbb{A})^{-1}v &= e^T \mathbb{B}^{-1}(\mathbb{U} - \mathbb{I})\Pi(\mathbb{I} - \mathbb{U}\Pi)^{-1}(\mathbb{U} - \mathbb{I})e \\ &= e^T \mathbb{B}^{-1}(\Pi - \mathbb{U})e + e^T \mathbb{B}^{-1}(\mathbb{I} - \Pi)(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e \end{aligned}$$

и

$$\begin{aligned} M(\mathbb{U}) &= \exp \left\{ \lambda v^T e + \lambda v^T \mathbb{A}[\mathbb{I} - \mathbf{V}\mathbb{A}]^{-1}v \right\} \\ &= \exp \left\{ \lambda [e^T \mathbb{B}^{-1}(\Pi - \mathbb{I})e + e^T \mathbb{B}^{-1}(\mathbb{I} - \Pi)(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e] \right\}. \end{aligned} \quad (17)$$

Согласно (12), $\ln M(\mathbb{U}) = \lambda [e^T \mathbb{B}^{-1}(\mathbb{I} - \Pi)(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e - 1]$. Матрица Π положительная, так же как и векторы $d = (\mathbb{I} - \Pi)e$ и $c = (\mathbb{I} - \Pi^T)\mathbb{B}^{-1}e$. Таким образом, (13) отвечает сложному закону Пуассона, т.е. $\ln M(\mathbb{U}) = \lambda [\Phi(\mathbb{U}) - 1]$, где $\Phi(\mathbb{U}) = c^T(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}d = \sum_{k=0}^{\infty} c^T(\mathbb{U}\Pi)^k \mathbb{U}d$ является производящей функцией усредняющего вероятностного распределения. Теорема 1 доказана.

4. Обсуждение. При $\Omega = \{i\}$ теорема 1 была доказана для независимых и одинаково распределенных величин в [3]. Условие $\mathbb{C}_n(1) \rightarrow \mu = 1/\beta_i > 1$, фигурирующее в этой статье, означает, что $\Pi = 1 - \beta_i$, так что (10) выполнено автоматически. Схожий результат имеется в [19], и он применялся в [15] для тестирования случайности. Существование предела в (9) является классической формулировкой редкости событий, вероятности которых имеют порядок $O(n^{-1})$ или длины которых ведут себя как $\ln n$.

Условие (11) означает, что для сужения $\mathbb{C}_m(1)$ матрицы $\mathbb{C}(1)$ из (2) на Ω_m

$$\lim_{n \rightarrow \infty} \mathbb{C}_m(1) = (\mathbb{I} - \Pi)^{-1}.$$

Производящая функция вероятностей усредняющего дискретного распределения, которое задает многомерный закон Пойа–Эпли в (13), может быть записана в виде

$$\Phi(\mathbb{U}) = c^T(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}d.$$

Средним этого распределения является вектор $\mathbb{B}^{-1}e$, а ковариационная матрица имеет вид

$$\mathbb{G} = \mathbb{B}^{-1}(\mathbb{I} - \Pi)^{-1} + (\mathbb{I} - \Pi^T)^{-1}\mathbb{B}^{-1} + \mathbb{B}^{-1} - \mathbb{B}^{-1}ee^T\mathbb{B}^{-1}.$$

Для одноточечного множества Ω усредняющее распределение — это геометрический закон на натуральных числах, т.е.

$$\Phi(u) = \frac{u(1 - \pi)}{1 - u\pi}. \quad (18)$$

В этой ситуации $M(u)$ отвечает классическому закону Пойа–Эпли.

Так же как и одномерном случае, случайный вектор, распределенный по закону Пойа–Эпли, может быть представлен в виде $\sum_{j=1}^M Y_j$, где случайная величина M , распределенная по закону Пуассона с параметром λ , не зависит от случайных векторов Y_j с усредняющим распределением. Другими словами, число M кластеров распределено по закону Пуассона, а их размеры Y_j имеют усредняющее распределение.

Имеем

$$\mathbf{E} \omega_i = \frac{\lambda}{\beta_i},$$

а ковариационная матрица распределения частот есть $\lambda\mathbb{G}$.

Отметим, что p -мерный закон Пойа–Эпли может быть также представлен в виде

$$\sum_{\mathbf{k} \geq \mathbf{0}} P_{\mathbf{k}} \mathbf{k}, \quad \mathbf{k}^T = (k_1, \dots, k_p), \quad (19)$$

где $P_{\mathbf{k}}$ — независимые случайные величины с пуассоновским распределением и средними $\lambda\gamma_{\mathbf{k}}$. Коэффициенты $\gamma_{\mathbf{k}}$ могут быть получены из разложения в степенной ряд $\Phi(U)$, т.е. являются вероятностями усредняющего закона. Ясно, что $\gamma_{0, \dots, 1, \dots, 0} = c_j d_j$. В общем случае эти вероятности могут быть найдены из многомерной формулы Лагранжа [7, п. 1.2.9]. Пусть $\text{Adj}(\mathbb{I} - \mathbb{U}\Pi)$ обозначает присоединенную к $\mathbb{I} - \mathbb{U}\Pi$ матрицу, тогда

$$\frac{c^T \text{Adj}(\mathbb{I} - \mathbb{U}\Pi) \mathbb{U}d}{\det(\mathbb{I} - \mathbb{U}\Pi)} = \Phi(\mathbb{U}) = \sum_{\mathbf{k} \geq \mathbf{0}} \gamma_{\mathbf{k}} u_1^{k_1} \dots u_p^{k_p}.$$

Согласно упомянутой теореме Лагранжа (см. также формулу (3) в [6]), если диагональная матрица $\mathbb{W}$ образована элементами $x_i/(e_i^T \Pi x)$, то $\gamma_{\mathbf{k}}$ является коэффициентом при соответствующей степени x степенного разложения функции

$$F_{\mathbf{k}}(x) = c^T \text{Adj}(\mathbb{I} - \mathbb{W}\Pi) \mathbb{W}d \prod_{i=1}^p (e_i^T \Pi x)^{k_i},$$

где $x = (x_1, \dots, x_p)^T$. Существование такого разложения предполагается.

Например, если $\Pi = ab^T$ для двух положительных векторов a, b таких, что выполнено условие (12), то

$$\begin{aligned} F_{\mathbf{k}}(x) &= \prod a_i^{k_i} (b^T x)^{\sum_i k_i} c^T \text{Adj} \left(\mathbb{I} - \frac{xb^T}{b^T x} \right) \mathbb{W} d \\ &= \prod a_i^{k_i} (b^T x)^{\sum_i k_i - 2} (c^T x) \left(\sum a_i^{-1} b_i d_i x_i \right), \end{aligned}$$

так как $\text{Adj}(\mathbb{I} - xb^T/(b^T x)) = xb^T/(b^T x)$. Таким образом, если $\sum k_i \geq 2$, то

$$\begin{aligned} \gamma_{\mathbf{k}} &= \sum_j \binom{\sum k_i - 2}{k_1 \dots k_j - 2 \dots k_p} (a_1 b_1)^{k_1} \dots (a_j b_j)^{k_j - 1} \dots (a_p b_p)^{k_p} c_j d_j \\ &\quad + \sum_{j < \ell} \binom{\sum k_i - 2}{k_1 \dots k_j - 1 \dots k_\ell - 1 \dots k_p} \\ &\quad \times (a_1 b_1)^{k_1} \dots (a_j b_j)^{k_j - 1} \dots (a_\ell b_\ell)^{k_\ell - 1} \dots (a_p b_p)^{k_p} (a_j c_j b_\ell d_\ell + a_\ell c_\ell b_j d_j). \end{aligned}$$

Параметры λ и β_i определены с точностью до общего множителя. Чтобы определить их однозначно, можно использовать условие нормировки (12), означающее, что $\Phi(0) = 0$, или что носитель усредняющего распределения не содержит нуля. Это условие может нарушаться для маргинальных распределений усредняющего закона, хотя из доказательства (17) видно, что такое представление имеет место для подмножеств Ω_m . Если $\mathbb{U}_1$ является подвектором $\mathbb{U}$, отвечающим словам из такого подмножества, то производящая функция $\Phi_1(\mathbb{U}_1)$ этих компонент вектора ω есть $\Phi(\mathbb{U}_1, \mathbb{I})$. Запишем $\Pi = \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{pmatrix}$ и соответственно представим векторы c и d .

Положим $\Pi_1 = \Pi_{11} + \Pi_{12}(\mathbb{I} - \Pi_{22})^{-1}\Pi_{21}$, так что $\Pi_1 = \mathbb{I} - [\mathbb{I} + (\Pi(\mathbb{I} - \Pi)^{-1})_{11}]^{-1}$. Имеем для подходящих $\tilde{c}, \tilde{d}$

$$\begin{aligned} \Phi_1(\mathbb{U}_1) &= c_2^T (\mathbb{I} - \Pi_{22})^{-1} d_2 + (c_1 - \Pi_{21}^T (\mathbb{I} - \Pi_{22}^T)^{-1} c_2)^T \\ &\quad \times (\mathbb{I} - \mathbb{U}_1 \Pi_1)^{-1} \mathbb{U}_1 (d_1 - \Pi_{12}^T (\mathbb{I} - \Pi_{22}^T)^{-1} d_2) \\ &= 1 + \tilde{c}^T [(\mathbb{I} - \mathbb{U}_1 \Pi_1)^{-1} \mathbb{U}_1 - (\mathbb{I} - \Pi_1)^{-1}] \tilde{d}. \end{aligned}$$

Получаем, что каждое одномерное маргинальное распределение является смесью геометрического закона на натуральных числах и вырожденного распределения в нуле:

$$\Phi(u_j) = \frac{u_j(1 - \pi_j)}{1 - u_j \pi_j} + 1 - c_j = \frac{1 - c_j + u_j(c_j - \pi_j)}{1 - u_j \pi_j},$$

где $0 \leq c_j \leq 1$, $0 \leq \pi_j \leq 1$.

Так как сумма усредняющих законов не обязана быть геометрическим законом, то асимптотическое распределение суммарной частоты $S = \sum_{i \in \Omega} \omega_i$ может и не иметь одномерных маргинальных распределений Пойа–Эпли, в то время как совместное распределение есть закон Пойа–Эпли. Однако представление (18) для S справедливо с $\pi = b^T a$, если $\Pi = ab^T$, где векторы a, b такие же, как и ранее. Теорема 1 верна и при более слабом условии, чем (12), а именно когда $c^T e \leq 1$. В этом случае усредняющее распределение задается производящей функцией

$$c^T(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e + 1 - c^T e. \quad (20)$$

Доказательство теоремы 1 показывает, что распределение всех частот $\omega_i, i \in \mathbb{Q}^\infty$, слабо сходится к процессу Пойа–Энли. Этот процесс задается счетной субстохастической матрицей Π_∞ такой, что оператор $\mathbb{I} - \Pi_\infty$ допускает обратный в пространстве всех ограниченных последовательностей. Также предполагается, что вектор $d = (\mathbb{I} - \Pi_\infty)e$ неотрицателен и c — вероятностный вектор. Конечномерные распределения, отвечающие словам в Ω_m , имеют производящие функции вида (20) для проекций c и d на соответствующее подпространство, а $\Pi = \mathbb{I} - [\mathbb{I} + \Pi_\infty(\mathbb{I} - \Pi_\infty)^{-1}|_{\Omega_m}]^{-1}$.

Если $\Pi = w\Xi$ для матрицы переходных вероятностей Ξ и числа w , $0 < w < 1$, то при $d = (\mathbb{I} - \Pi)e = (1 - w)e$

$$\begin{aligned} \Phi(\mathbb{U}) &= c^T(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}d = (1 - w) \sum_{N=1}^{\infty} w^{N-1} c^T(\mathbb{U}\Pi)^{N-1}\mathbb{U}e \\ &= (1 - w) \sum_{N=1}^{\infty} w^{N-1} \Phi_N(\mathbb{U}), \end{aligned}$$

где $\Phi_N(\mathbb{U}) = c^T(\mathbb{U}\Pi)^{N-1}\mathbb{U}e = c^T\mathbb{U}(\Pi\mathbb{U})^{N-1}e$ — производящая функция в (8). Эта функция отвечает вектору частот марковской цепи длины N , переходные вероятности которой задаются матрицей Ξ , а c есть вектор начальных состояний. Иными словами, усредняющее распределение является смесью распределений частот в марковских цепях случайной геометрической длины.

Если Π — диагональная матрица, то $\ln M(\mathbb{U})$ есть сумма $\ln M(u_j)$ из (18) с параметрами $\lambda_j = \lambda c_j d_j$ и $\pi_j = \Pi_{jj}$ при условии, что $\sum_j c_j d_j (1 - \Pi_{jj})^{-1} = 1$. В частности, если Ω_m состоит из апериодических некоррелированных слов i^m таких, что $\mathbb{C}_{i^m i^m}(1) = 1$ и $\mathbb{C}_{i^m j^m}(1) = 0$, $i^m \neq j^m$, то $\Pi = 0$ и предельным распределением $\omega_i, i \in \Omega$, является произведение классических законов Пуассона. Если Π является блочно-диагональной матрицей, то ω_i и ω_j для i и j из разных блоков независимы. Обратное также верно.

В [1] были получены верхние оценки расстояния по вариации между сложным законом Пуассона и распределением частот редких слов. Хотелось бы надеяться на справедливость усиленного варианта теоремы 1,

предоставляющего точную оценку такого расстояния в многомерном случае. Возможно, что метод Стейна с применением (19) окажется здесь полезным [2]. Однако эта идея наталкивается на значительные трудности. Известные результаты в этом направлении [11], [4], [5], [14] далеки от оценок, которые бы могли доказать сходимость к закону Пойа–Эпли при наших условиях (9)–(11).

5. Примеры. Мы начнем со случая независимых и равномерно распределенных $\varepsilon_1, \dots, \varepsilon_n$, когда $q = 2$. Предположим, что $n \rightarrow \infty$, $n/2^m \rightarrow \lambda/\beta$, так что $\beta_i \equiv \beta$. Эта модель отвечает задаче о проверке случайности. Условие (12) означает, что $\beta = e^T(\Pi - \Pi)e$. Для $m = 2k$ определим три цепочки $i^m = (1, 0, 1, 0, \dots, 1, 0)$, $j^m = (0, 1, 0, 1, \dots, 0, 1)$ и $\ell^m = (1, 1, 1, \dots, 1)$. Тогда

$$\begin{aligned}\mathbb{C}_{i^m i^m}(z) &= \sum_{r=0}^k \frac{z^{2r}}{2^{2r}} = \mathbb{C}_{j^m j^m}(z), & \mathbb{C}_{\ell^m \ell^m}(z) &= \sum_{r=0}^m \frac{z^r}{2^r}, \\ \mathbb{C}_{i^m j^m}(z) &= \mathbb{C}_{j^m i^m}(z) = \sum_{r=1}^k \frac{z^{2r-1}}{2^{2r-1}}, \\ \mathbb{C}_{i^m \ell^m}(z) &= \mathbb{C}_{j^m \ell^m}(z) = \mathbb{C}_{\ell^m j^m}(z) = \mathbb{C}_{\ell^m i^m}(z) = 0.\end{aligned}$$

Поэтому

$$\mathbb{C}_m - \mathbb{I} \rightarrow \frac{1}{3} \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \quad \Pi = \frac{1}{2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$\beta = 3/2$. Согласно теореме 1,

$$\mathbf{E} u_1^{\omega_i} u_2^{\omega_j} u_3^{\omega_\ell} \rightarrow \exp \left\{ \frac{2\lambda}{3} \left[\frac{u_1 + u_2 + 2u_1 u_2 - 4}{4 - u_1 u_2} + \frac{u_3 - 1}{2 - u_3} \right] \right\}.$$

Таким образом, цепочки i, j асимптотически не зависят от цепочки ℓ (распределение которой есть классический закон Пойа–Эпли).

Если же $i^m = (0, 0, 1, \dots, 1, 1)$, а $j^m = (0, 1, 1, \dots, 1, 1)$, то для той же цепочки ℓ

$$\Pi = \frac{1}{2} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

и $\beta = 3/2$. Таким образом,

$$\mathbf{E} u_1^{\omega_i} u_2^{\omega_j} u_3^{\omega_\ell} \rightarrow \exp \left\{ \frac{\lambda}{6} \left(\frac{u_1 u_2 u_3 + u_2 u_3}{2 - u_3} + 2u_1 + u_2 + u_1 u_2 - 6 \right) \right\}.$$

Если $\kappa_1^p = (0, 1, \dots, 1)$, $\kappa_2^p = (1, \dots, 1, 0)$ — два слова длины p , а $i^{rm} = (\kappa_1, \dots, \kappa_1)$, $j^{rm} = (\kappa_2, \dots, \kappa_2)$ — их повторения r раз, так что

$m = pr$, то для $\Omega = \{i, j\}$

$$\mathbb{C}_m(1) - \mathbb{I} \rightarrow \frac{1}{2^p - 1} \begin{pmatrix} 1 & 2^{p-1} \\ 2 & 1 \end{pmatrix}.$$

Таким образом,

$$\Pi = \begin{pmatrix} 0 & \frac{1}{2} \\ 2^{-(p-1)} & 0 \end{pmatrix}.$$

Получаем $\beta = 3/2 - 1/2^{p-1}$ и

$$\mathbf{E} u_1^{\omega_i} u_2^{\omega_j} \rightarrow \exp \left\{ \lambda \left[\frac{2^{p-1} - 1}{\beta(2^p - u_1 u_2)} \left(\frac{(2^{p-1} - \beta)u_1 u_2}{2^{p-1} - 1} + u_1 + u_2 \right) - 1 \right] \right\}.$$

Эти примеры наводят на мысль о новом варианте теста случайности на основе частот перекрывающихся слов [15]. Цепочки $i^m = (1, 0, 1, 0, \dots, 1, 0)$ и ℓ^m могут послужить основой для такого критерия.

Перейдем теперь к цепям Маркова, скажем, порядка один. В анализе ДНК $q = 4$; пусть Ω составляет мотив, т.е. $\Omega_m = \{(i_1, \dots, x, \dots, i_m), x = a, c, g, t\}$ для заданных $i_1, \dots, i_{s-1}, i_{s+1}, \dots, i_m$ и $p = q$. Предположим, что матрица переходных вероятностей $\Xi = \Xi_m$ известна (на практике она может быть надежно оценена). Пусть вектор ξ является начальным и стационарным распределением, $\Xi^T \xi = \xi$.

Доопредельный вариант условия (11) можно проверять, основываясь на оценках матрицы $\mathbb{C}_m$ и пользуясь следующей формулой [17, формула (10)] для элементов $\mathbb{C}_m(1)$:

$$\begin{aligned} \mathbb{C}(i^m, j^m) &= \delta_{i^m j^m} + \sum_{r=1}^{m-1} \delta_{(i_{r+1} \dots i_m), (j_1 \dots j_{m-r})} \xi_{j_{m-r} j_{m-r+1}} \dots \xi_{j_{m-1} j_m} \\ &\quad + \xi_{j_1 j_2} \dots \xi_{j_{m-1} j_m} [\Xi(\mathbb{I} - \Xi + e\xi^T)^{-1}]_{i_m j_1} - P(j^m). \end{aligned}$$

Если матрица $\mathbb{K}_m$ задана элементами

$$\sum_{r=1}^{m-1} \delta_{(i_{r+1} \dots i_m), (j_1 \dots j_{m-r})} \xi_{j_{m-r} j_{m-r+1}} \dots \xi_{j_{m-1} j_m},$$

то

$$\lim_{m \rightarrow \infty} \mathbb{C}_m(1) = \lim_{m \rightarrow \infty} (\mathbb{I} + \mathbb{K}_m).$$

В самом деле,

$$\begin{aligned} &\xi_{j_1 j_2} \dots \xi_{j_{m-1} j_m} [\Xi(\mathbb{I} - \Xi + e\xi^T)^{-1}]_{i_m j_1} - P(j^m) \\ &= \xi_{j_1 j_2} \dots \xi_{j_{m-1} j_m} \sum_k \xi_{i_m k} \left[(\mathbb{I} - \Xi + e\xi^T)_{k j_1}^{-1} - \xi_{j_1} \right], \end{aligned}$$

так что для $\zeta = \max \xi_{rs} < 1$ имеем

$$\max_{jk} |\mathbb{C}_m(1) - \mathbb{I} - \mathbb{K}_m|_{jk} \leq \zeta^{m-1} \max_{jk} |(\mathbb{I} - \Xi + e\xi^T)_{kj}^{-1} - \xi_j| \rightarrow 0.$$

Тем самым, достаточно проверять (11) для матрицы $\mathbb{I} + \mathbb{K}_m$. Если это условие выполнено и последовательность такова, что $nP(i) = n\xi_{i_1}\xi_{i_1i_2}\cdots\xi_{i_{m-1}i_m}$ не слишком мало, то можно положить $p = \{P(i)\}$,

$$\beta_i^{-1} = \frac{P(i)}{e^T \mathbb{C}_m^{-1} p}, \quad \lambda = \frac{n \sum \beta_i P(i)}{q}.$$

В этой ситуации многомерный закон Пойа–Эпли может служить для идентификации участков с аномально редкими или частыми словами.

Прямые вычисления в примере раздела 7.1 из [13] дают $m = 8$, $i_1 = i_4 = i_5 = i_7 = i_8 = g$, $i_3 = i_6 = t$, $s = 2$, $n = 1\,830\,140$,

$$\begin{aligned} \lambda &= 55.87, \\ 1/\beta &= [0.2670, 0.2209, 0.1749, 0.3456]^T, \\ c &= [0.2679, 0.2217, 0.1637, 0.3467]^T. \end{aligned}$$

Имеем $d = 0.9881e$ и для $a = [0, 0, 0.0117, 0.0001]^T$

$$\Pi = ea^T.$$

Метод, обсуждавшийся в предыдущем пункте, позволяет найти вероятности $\gamma_{\mathbf{k}}$ усредняющего распределения, основная масса которого (98.46%) сосредоточена на плоскости $\sum k_i = 1$:

$$\gamma_{1,0,0,0} = 0.2638, \quad \gamma_{0,1,0,0} = 0.2183, \quad \gamma_{0,0,1,0} = 0.1611, \quad \gamma_{0,0,0,1} = 0.3414.$$

Математические ожидания частот слов из этого мотива совпадают с ответом, который предлагает метод работы [13]. Однако наш метод также выдает ковариационную матрицу частот. Обычная корреляционная матрица имеет вид

$$\begin{pmatrix} 1.0000 & -0.3212 & -0.2553 & -0.4384 \\ -0.3212 & 1.0000 & -0.2253 & -0.3868 \\ -0.2553 & -0.2253 & 1.0000 & -0.3074 \\ -0.4384 & -0.3868 & -0.3074 & 1.0000 \end{pmatrix}.$$

Аналогичные вычисления можно проводить и в значительно более сложных ситуациях, например когда Ω_m содержит коррелированные цепочки.

Автор благодарен профессору J. Lawrence за приведенное короткое доказательство формулы (16) и рецензенту за полезные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. Barbour A. D., Chen L. H. Y., Loh W.-L. Compound Poisson approximation for non-negative random variables via Stein's method. — Ann. Probab., 1992, v. 20, № 4, p. 1843–1866.
2. Barbour A. D., Chrysaphinou O. Compound Poisson approximation: a user's guide. — Ann. Appl. Probab., 2001, v. 11, № 3, p. 964–1002.
3. Chrysaphinou O., Papastavridis S. A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials. — Probab. Theory Related Fields, 1988, v. 79, № 1, p. 129–143.
4. Erhardsson T. Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. — Ann. Appl. Probab., 2000, v. 10, № 2, p. 573–591.
5. Erhardsson T. Stein's method for Poisson and compound Poisson approximation. — An Introduction to Stein's method. Ed. by A. Barbour and L. H. Y. Chen. Singapore: Singapore Univ. Press, 2005, p. 61–113.
6. Good I. J. The frequency count of a Markov chain and the transition to continuous time. — Ann. Math. Statist., 1961, v. 32, p. 41–48.
7. Гульден Я., Джексон Д. Перечислительная комбинаторика. М.: Наука, 1990, 503 с.
8. Guibas L. J., Odlyzko A. M. Long repetitive patterns in random sequences. — Z. Wahrscheinlichkeitstheor. verw. Geb., 1980, v. 53, № 3, p. 241–262.
9. Guibas L. J., Odlyzko A. M. Strings overlaps, pattern matching and nontransitive games. — J. Combin. Theory Ser. A, 1970, v. 30, № 2, p. 183–208.
10. Кемени Дж. Дж., Снелл Дж. Л. Конечные цепи Маркова. М.: Наука, 1970, 271 с.
11. Reinert G., Schbath S. Large compound Poisson approximations for occurrences of multiple words. — Statistics in Molecular Biology and Genetics (Seattle, 1997). Hayward: Inst. Math. Statist., 1999, p. 257–275. (IMS Lecture Notes Monogr. Ser., v. 33.)
12. Риордан Дж. Введение в комбинаторный анализ. М.: ИЛ, 1963, 287 с.
13. Robin S., Rodolphe F., Schbath S. DNA, Words and Models: Statistics of Exceptional Words. Cambridge: Cambridge Univ. Press, 2006.
14. Roquain E., Schbath S. Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain. — Adv. Appl. Probab., 2007, v. 39, № 1, p. 128–140.
15. Rukhin A. L. Testing randomness: a suite of statistical procedures. — Теория вероятн. и ее примен., 2000, т. 45, в. 1, с. 111–132.
16. Rukhin A. L. Distribution of the number of words with a prescribed frequency and tests of randomness. — Adv. Appl. Probab., 2002, v. 34, № 4, p. 775–797.
17. Рухин А. Л. Корреляционные матрицы цепочек для марковских последовательностей и тестирование случайности. — Теория вероятн. и ее примен., 2006, т. 51, в. 4, с. 712–731.
18. Szpankowski W. Average Case Analysis of Algorithms on Sequences. New York: Wiley-Interscience, 2001, 551 p.
19. Зубков А. М., Михайлов В. Г. Предельные распределения случайных величин, связанных с длинными повторениями в последовательности независимых испытаний. — Теория вероятн. и ее примен., 1974, т. 19, в. 1, с. 173–181.
20. Зубков А. М., Михайлов В. Г. О повторениях s -цепочек в последовательности независимых величин. — Теория вероятн. и ее примен., 1979, т. 24, в. 2, с. 269–282.

Поступила в редакцию
6.X.2007

Исправленный вариант
30.XII.2008