THEORY PROBAB. APPL.
Vol. 54, No. 2, pp. 246–260

Translated from Russian Journal

JOINT DISTRIBUTION OF PATTERN FREQUENCIES AND MULTIVARIATE PÒLYA–AEPPLI LAW*

A. L. RUKHIN†

(*Translated by the author*)

Abstract. The paper studies the joint distribution of frequencies of overlapping words in a Markov sequence. Usually characteristics of this distribution are expressed in terms of a so-called pattern correlation matrix. A more direct approach allows for explicit formulas which involve the fundamental matrix of the Markov chain whose states are words of a given length. These formulas lead to the probability generating function of the asymptotic joint distribution of pattern frequencies corresponding to a new multivariate discrete distribution.

Key words. Bell polynomials, compound Poisson distribution, fundamental matrix, Lagrange formula, Markov chains, pattern correlation matrix, Pòlya–Aeppli distribution, successions

DOI. 10.1137/S0040585X97984115

1. Introduction. Consider a random text formed by realizations of letters chosen from a finite alphabet. For a given set of patterns (or words) it is of interest to determine the probability of the prescribed number of their occurrences in the text. This problem appears in different areas of information theory (source coding, randomness testing) and in molecular biology (DNA analysis, gene recognition). Chapter 7 of [18] reviews some information-theoretic aspects. Applications of occurrence counts distributions in DNA analysis are discussed at an elementary level in [13]. Under an independence assumption, the formulas for such probabilities depend on the so-called correlation polynomial introduced by Guibas and Odlyzko [9].

In section 2 we discuss the pattern correlation matrix definition for a regular Markov chain and look at its relationship to the fundamental matrix of the Markov chain whose states are the patterns of the given length. The form of the probability generating function of word frequencies is derived. It leads to the asymptotic distribution of these frequencies which is shown to be a special multivariate compound Poisson law in section 3. That law, the multivariate Pòlya–Aeppli distribution, is discussed in section 4. The concluding section 5 contains some examples.

The mentioned asymptotic distribution is derived under the assumption that the pattern length m as well as the order of the Markov chain increase to infinity. A similar condition has been used in [19], [20], [8], and [3].

2. Pattern correlation matrices. Let a random text $X_n = (\varepsilon_1, \dots, \varepsilon_n)$ be formed by a stationary series of discrete random variables ε_k taking possible values in the alphabet $\mathbb{Q} = \{1, \dots, q\}$ for some positive integer $q \geq 2$. We start with a Markov chain of fixed order $m, m \geq 1$. Many statistical procedures designed to analyze such a series are based on the observed counts of overlapping m-patterns (words or templates), like $\imath^m = (i_1 \dots i_m)$.

*Received by the editors October 6, 2007; revised December 30, 2008. This work was supported by NSA grant H98230-06-1-0068.

http://www.siam.org/journals/tvp/54-2/98411.html

†National Institute of Standards and Technology SED, Gaithersburg, MD 20899, and Department of Mathematics and Statistics UMBC, 1000 Hilltop Circle, Baltimore, MD 21250 (andrew.rukhin@nist.gov).

Denote the set of all these patterns (of cardinality q^m) by $\mathbb{Q}^m$. It will be assumed that for $\imath = \imath^m \in \mathbb{Q}^m$, the probabilities $P(\imath) = \mathbf{P}((\varepsilon_1, \dots, \varepsilon_m) = (i_1 \dots i_m))$ are positive. Let $P(j_{k+1} \cdots j_m | \imath) = P(\imath j_{k+1} \dots j_m)/P(\imath)$ denote the conditional probability of the word $j_{k+1} \dots j_m$, $k = 1, \dots, m$, preceded by the pattern $\imath$. Also denote by $\mathbb{I}$ the identity matrix, by e the q^m-dimensional vector with all coordinates equal to one, by p the q^m-dimensional vector with coordinates $P(\imath), \imath \in \mathbb{Q}^m$, and let the $q^m \times q^m$ matrix $\mathbb{P} = \mathbb{P}_m$ be formed by the entries

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

Here and below $\delta_{\ell,\kappa} = \prod_{k=1}^{K} \delta_{l_k \kappa_k}$ is the Kronecker symbol for two K-indices $\ell = (\ell_1, \dots, \ell_K)$ and $\kappa = (\kappa_1, \dots, \kappa_K)$.

The pattern correlation matrix $\mathbb{C}(z)$ can be defined for all complex z, for which $[\mathbb{I} - z(\mathbb{P} - ep^T)]^{-1}$ exists:

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

This matrix has been used in [18] for Bernoulli random sequences. It leads to the form of generating functions for the probabilities of joint pattern occurrences in a random text [16], [17].

The pattern correlation polynomial, $C_{\imath\jmath}(z)$, was introduced for words $\imath = (i_1 \dots i_m)$ and $\jmath = (j_1 \dots j_m)$ in [9] as

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

With $C(z)$ denoting the matrix formed by the pattern correlation polynomials $C_{\imath\jmath}(z)$, one gets for $0 \le z < 1$ the following formula for the pattern correlation matrix,

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

as $\mathbb{P}^T p = p$, and $\mathbb{P}e = e$. Notice that $\mathbb{C}(z) = C(z)$ in the case of independent and identically distributed variables. In this situation all eigenvalues of $\mathbb{P}$ are 1 (with multiplicity one) or 0 (with multiplicity $q^m - 1$), so that $\mathbb{C}(z)$ is defined for all z. According to the definition,

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

While the pattern correlation polynomials $C_{\imath\jmath}(z)$ suffice for the pattern analysis for independent and identically distributed random variables, the general case of Markov dependence demands the pattern correlation matrix. Indeed the matrix $\mathbb{C}(1)$ plays a crucial role in the Markov process Y_n of order one, whose states are words of length m, i.e., $Y_n = \imath$ if $(\varepsilon_n, \dots, \varepsilon_{n+m-1}) = (i_1 \dots i_m)$. For this process the transition probabilities have the form (1). We assume that the transition matrix $\mathbb{P}$ has a unique eigenvalue of modulus 1, so that the chain is ergodic or regular. This condition holds

if all probabilities $P(\imath^{m+1})$ are positive, and then $\rho(\mathbb{P}-ep^T)<1$ with $\rho(A)$ denoting the spectral radius of a matrix A. The frequency, $\omega_\imath=\omega_\imath(n)$, of the pattern $\imath$ in the original sequence equals the number of occurrences of the state $\imath$ for the process Y_n.

It is well known that the asymptotic joint distribution of these frequencies is normal. For our matrix $\mathbb{P}$, the probability vector of the stationary distribution is p, i.e., is formed by coordinates $P(\imath)=P(\varepsilon_1=i_1,\dots,\varepsilon_m=i_m)$. The fundamental matrix $\mathbb{Z}$ of the Markov chain Y_n is defined by the formula, $\mathbb{Z}=(\mathbb{I}-\mathbb{P}+ep^T)^{-1}$. It admits a simple expression through $\mathbb{C}(1)$,

$$\mathbb{Z}=\mathbb{C}(1)-(m-1)ep^T. \tag{4}$$

An important property of the fundamental matrix is that it determines the limiting covariance matrix of the frequencies distribution. According to Theorem 4.6.1 in [10] when m is fixed, the limiting covariance matrix $\Sigma=\lim_{n\to\infty}n^{-1}\operatorname{cov}(\omega_\imath,\omega_\jmath)$ is formed by the elements

$$\sigma_{\imath\jmath}=-\delta_{\imath\jmath}P(\imath)-P(\imath)P(\jmath)+P(\imath)\mathbb{Z}_{\imath\jmath}+P(\jmath)\mathbb{Z}_{\jmath\imath}. \tag{5}$$

Because of (4), Σ can be written as

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

where $\mathbb{D}=\operatorname{diag}(p)$ is a diagonal matrix.

Thus the limiting covariance matrix of the joint distribution of empirical frequencies of all m-patterns in a random, order m Markov text admits a simple expression through the pattern correlation matrix $\mathbb{C}$.

3. Probability generating functions of frequencies and their asymptotic behavior. We start with the following result which gives a convenient representation of the probability generating function of frequencies $\omega_\imath(n)$, $\imath\in\mathbb{Q}^m$.

LEMMA 1. *Let $X_n=(\varepsilon_1,\dots,\varepsilon_n)$ be a Markov chain of order m with $\omega_\imath(n)$, $\imath\in\mathbb{Q}^m$, denoting the frequencies of overlapping m-words, $\sum_\imath\omega_\imath=n-m+1=N$. The joint probability generating function of these frequencies has the form*

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

where the sum with regard to $\ell_0,\dots,\ell_s$ is taken over the set $\mathbb{Q}^{m(s+1)}$, e_ℓ denotes the ℓth basis vector, $\ell\in\mathbb{Q}^m$, and

$$\mathbb{T}_N(s)=\sum_{\substack{i_1,\dots,i_s\geqq 1\\ i_1+\cdots+i_s\leqq N}}(N-i_1-\cdots-i_s)\mathbb{P}^{i_1}\otimes\cdots\otimes\mathbb{P}^{i_s}, \tag{7}$$

with the matrix $\mathbb{P}$ defined in (1), *and $\mathbb{P}_1\otimes\mathbb{P}_2$ denoting the tensor product of the matrices* (*or vectors*) *$\mathbb{P}_1$ and $\mathbb{P}_2$.*

Proof. To derive this representation, we use a known formula for generating functions of frequencies in Markov chains (see [6, p. 41]), according to which, with $\mathbb{U}$ denoting the diagonal matrix formed by elements $u_\imath$, $\imath\in\mathbb{Q}^m$,

$$M_N(\mathbb{U})=\mathbf{E}\prod_\imath u_\imath^{\omega_\imath}=p^T\mathbb{U}(\mathbb{P}\mathbb{U})^{N-1}e. \tag{8}$$

One has with $\tilde{\mathbb{P}} = \mathbb{P}(\mathbb{U} - \mathbb{I})$,

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

so that the terms in the sum over $i_1, \dots, i_{s+1}$ have the matrix $\widetilde{\mathbb{P}}$ entering exactly s times. Therefore, as $\mathbb{P}^{i_{s+1}}e = e$, and $\widetilde{\mathbb{P}}e = \sum_\ell (u_\ell - 1)\mathbb{P}e_\ell$,

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

Since $p^T\mathbb{P}^{i_1}\widetilde{\mathbb{P}} = p^T\widetilde{\mathbb{P}} = p^T(\mathbb{U} - \mathbb{I})$,

$$\begin{aligned}p^T(\mathbb{P}\mathbb{U})^{N-1}e &= 1 + (N-1)\sum_\ell (u_\ell - 1)p^T e_\ell \\ &\quad + \sum_\ell (u_\ell - 1) \sum_{s=2}^{N-1} \sum_{\substack{i_2,\dots,i_s \geqq 0 \\ i_2+\cdots+i_s \leqq N-1-s}} (N - s - i_2 - \cdots - i_s)\, p^T\widetilde{\mathbb{P}}\mathbb{P}^{i_2}\widetilde{\mathbb{P}}\cdots\widetilde{\mathbb{P}}\mathbb{P}^{i_s+1}e_\ell \\ &= 1 + (N-1)\sum_\ell (u_\ell - 1)P(\ell) + \sum_{\jmath,\,\ell} (u_\jmath - 1)(u_\ell - 1)P(\jmath) \\ &\quad \times \sum_{s=1}^{N-2} \sum_{\substack{i_1,\dots,i_s \geqq 0 \\ i_1+\cdots+i_s \leqq N-2-s}} (N - 1 - s - i_1 - \cdots - i_s) e_\jmath^T \mathbb{P}^{i_1}\widetilde{\mathbb{P}}\cdots\widetilde{\mathbb{P}}\mathbb{P}^{i_s+1}e_\ell.\end{aligned}$$

Similarly,

$$\begin{aligned}p^T(\mathbb{U} - \mathbb{I})(\mathbb{P}\mathbb{U})^{N-1}e &= \sum_\ell (u_\ell - 1)P(\ell) \\ &\quad + \sum_{\jmath,\,\ell} (u_\jmath - 1)(u_\ell - 1)P(\jmath) \sum_{s=1}^{N-1} \sum_{\substack{i_1,\dots,i_s \geqq 0 \\ i_1+\cdots+i_s \leqq N-1-s}} e_\jmath^T \mathbb{P}^{i_1}\widetilde{\mathbb{P}}\cdots\widetilde{\mathbb{P}}\mathbb{P}^{i_s+1}e_\ell.\end{aligned}$$

Thus,

$$\begin{aligned}p^T\mathbb{U}(\mathbb{P}\mathbb{U})^{N-1}e &= 1 + N\sum_\ell (u_\ell - 1)P(\ell) + \sum_{\jmath,\,\ell} (u_\ell - 1)(u_\jmath - 1)P(\jmath) \\ &\quad \times \sum_{s=1}^{N-1} \sum_{\substack{i_1,\dots,i_s \geqq 0 \\ i_1+\cdots+i_s \leqq N-1-s}} (N - s - i_1 - \cdots - i_s) e_\jmath^T \mathbb{P}^{i_1}\widetilde{\mathbb{P}}\cdots\widetilde{\mathbb{P}}\mathbb{P}^{i_s+1}e_\ell.\end{aligned}$$

For $k = 1, \dots, s-1$, replace the kth factor $\widetilde{\mathbb{P}}$ in this formula by $\sum_{\ell_k}(u_{\ell_k}-1)\mathbb{P}e_{\ell_k}e_{\ell_k}^T$, put $\ell_0 = \jmath, \ell_s = \ell$, and plug in (8) to obtain

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

In our application of Lemma 1 the patterns belong to a given set $\Omega_m \subset \mathbb{Q}^m$, in which case the sum in (6) is taken over $\ell_0, \dots, \ell_s \in \Omega_m$.

Our main result (Theorem 1) refers to a sequence of m-patterns, and to the joint asymptotic distribution of their frequencies $\omega_{\imath^m} = \omega_{\imath^m}(n), \imath^m \in \mathbb{Q}^m$, when $n \to \infty, m = m(n) \to \infty$, so that $N/n \to 1$. In this theorem we assume that Ω is a finite set of fixed cardinality formed by infinite sequences $\imath = (\imath_1, \dots, \imath_m, \dots)$. The set $\Omega_m, m = 1, 2, \dots$, is a restriction of Ω, i.e., is composed by m-patterns, $\imath^m = (\imath_1, \dots, \imath_m)$.

To simplify notation, now let $\mathbb{P}$ be the submatrix of the matrix defined by (1) formed by rows and columns corresponding to patterns in Ω_m; $\mathbb{I}$ will be the identity matrix of the appropriate size, with a similar agreement for vectors e and p.

THEOREM 1. *Let $\{X_n = (\varepsilon_1, \dots, \varepsilon_n),\ n = 1, 2, \dots\}$ be a sequence of ergodic Markov chains with X_n of order $m = m(n)$. Assume that $n \to \infty$, so that $m \to \infty$, $m/n \to 0$, and for all $\imath \in \Omega$,*

$$\lim_{n\to\infty} nP(\imath^m) = \frac{\lambda}{\beta_\imath}, \tag{9}$$

with positive λ and $\beta_\imath$. Suppose further that

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

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

for some substochastic matrix Π such that with the diagonal matrix $\mathbb{B}$ formed by $\beta_\imath$, $\mathbb{B}\Pi^T\mathbb{B}^{-1}$ is also a substochastic matrix, and

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

Then the joint distribution of frequencies $\omega_{\imath^m}, \imath^m \in \Omega_m$, converges to the multivariate compound Poisson distribution (Pòlya–Aeppli law)

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

Proof. To employ Lemma 1 notice that for a given m, s, and $\ell_0, \dots, \ell_s \in \Omega$,

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

for any positive integers $i_1, \dots, i_s$, $i_1 + \cdots + i_s \le N - 1$, and

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

where $t_1, \dots, t_s$, $t_1 < \cdots < t_r$, is a permutation of integers $0, 1, \dots, s-1$. Because of (11), for any $\jmath$ and ℓ from Ω,

$$\sum_{i \geqq 1} e_{\ell^m}^T (\mathbb{P} - ep^T)^i e_{\jmath^m} = e_{\ell^m}^T (\mathbb{I} - \mathbb{P} + ep^T)^{-1} e_{\jmath^m} - \delta_{\ell^m \jmath^m} \to \frac{\alpha_{\ell\jmath}}{\beta_\jmath},$$

with $\alpha_{\ell\jmath}$ denoting entries of the matrix $\mathbb{A} = (\mathbb{I}-\Pi)^{-1}\mathbb{B} - \mathbb{B} = \Pi(\mathbb{I}-\Pi)^{-1}\mathbb{B}$. According to (10), as $i \to \infty$, $e_{\ell^m}^T (\mathbb{P} - ep^T)^i e_{\jmath^m}$ tends to zero uniformly in m, so that the series above converges uniformly in m.

For fixed s and $1 \leq r \leq s$,

$$\lim_{n\to\infty} \sum_{i_{t_1+1}, \dots, i_{t_r+1} \geqq 1} \prod_{k=1}^{r} e_{\ell^m_{t_k}}^T (\mathbb{P} - ep^T)^{i_{t_k+1}} e_{\ell^m_{t_k+1}}$$
$$= \lim_{n\to\infty} \prod_{k=1}^{r} \left(e_{\ell^m_{t_k}}^T (\mathbb{I} - \mathbb{P} + ep^T)^{-1} e_{\ell^m_{t_k+1}} - \delta_{\ell^m_{t_k} \ell^m_{t_k+1}} \right) = \frac{\alpha_{\ell_1\ell_2} \cdots \alpha_{\ell_r \ell_{r+1}}}{\beta_{\ell_2} \cdots \beta_{\ell_{r+1}}}.$$

Therefore, for a given s, with $\mathbb{T}_N(s)$ defined by (7),

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

where $\mathbb{S}_r$ is the rth elementary symmetric function,

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

Indeed, if $i_{u_1}, \dots, i_{u_{s-r}}$ denote indices different from $i_{t_1+1}, \dots, i_{t_r+1}$, then for $I = \sum(i_{t_1+1} + \cdots + i_{t_r+1})$,

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

and by uniform convergence of the series, $\sum_{i_{t_1+1},\dots,i_{t_r+1}\ge 1}\prod_{k=1}^r e_{\ell^m_{t_k}}^T(\mathbb{P}-ep^T)^{i_{t_k+1}}e_{\ell^m_{t_k+1}}$,

$$\lim_{n\to\infty}\frac{(s-r+1)!}{N^{s-r+1}}\sum_{\substack{i_{t_1+1},\dots,i_{t_r+1}\ge 1\\ i_{t_1+1}+\dots+i_{t_r+1}\le N-s+r-1}}\binom{N-i_{t_1+1}-\dots-i_{t_r+1}}{s-r+1}$$
$$\times\prod_{k=1}^r e_{\ell^m_{t_k}}^T(\mathbb{P}-ep^T)^{i_{t_k+1}}e_{\ell^m_{t_k+1}}$$
$$=\lim_{n\to\infty}\left[\binom{N}{s-r+1}\right]^{-1}\sum_{i=0}^{N-s}\binom{N-i-r}{s-r}$$
$$\times\sum_{i_{t_1}+\dots+i_{t_r}\le i+r-1}\prod_{k=1}^r e_{\ell^m_{t_k}}^T(\mathbb{P}-ep^T)^{i_{t_k+1}}\,e_{\ell^m_{t_k+1}}=\frac{\alpha_{\ell_{t_1}\ell_{t_1+1}}\cdots\alpha_{\ell_{t_r}\ell_{t_r+1}}}{\beta_{\ell_{t_1+1}}\cdots\beta_{\ell_{t_r+1}}}.$$

It follows from (6) that $\lim_{n\to\infty}(M_N(\mathbb{U})/M(\mathbb{U}))=1$, where with $v_\ell=\beta_\ell^{-1}(u_\ell-1)$,

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

In the sum in (14) $\ell_0,\dots,\ell_s\in\Omega$; to evaluate it, we partition the set $\{t_1<\dots<t_r\}$ according to the number of successions of given lengths. More precisely, let us call the string $t_{i_1}<\dots<t_{i_p}$ a *succession* of length p if $t_{i_{k+1}}=t_{i_k}+1$ for $k=1,\dots,p-1$, but $t_{i_{p+1}}\ne t_{i_p}+1$ and $t_{i_1}\ne t_{i_1-1}+1$. (If $t_1=0$ or $t_p=r$, these inequalities hold automatically.) Enumeration of strings with a given number of successions is discussed in [7, section 2.3.15].

The sums over the indices forming a succession of length p can be readily evaluated,

$$\sum_{t_{i_1}<\dots<t_{i_p}}\alpha_{\ell_{t_{i_1}}\ell_{t_{i_1}+1}}\alpha_{\ell_{t_{i_2}}\ell_{t_{i_2}+1}}\cdots\alpha_{\ell_{t_{i_p}}\ell_{t_{i_p}+1}}v_{\ell_{t_{i_1}}}v_{\ell_{t_{i_2}}}\cdots v_{\ell_{t_{i_p}}}v_{\ell_{t_{i_p}+1}}$$
$$=v^T\mathbb{A}(\mathbf{V}\mathbb{A})^{p-1}v=:z_p \tag{15}$$

where $\mathbb{V}$ is the diagonal matrix formed by the vector v with the coordinates $v_\ell=\beta_\ell^{-1}(u_\ell-1)$.

Denote for a fixed s by $C_s(\nu_1,\dots,\nu_r)$ the number of sequences, $t_1<\dots<t_r$, formed by integers $0,\dots,s-1$ with ν_p successions of length p, $p=1,\dots,r$. Clearly, $\sum p\nu_p=r$, and

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

We now prove that

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

Indeed, if $\sum\nu_p=h$, to construct the set of r numbers $t_1<\cdots<t_r$ from $\{0,1,\dots,s\}$, first choose $h+1$ positive integers $a_0,a_1,\dots,a_h$, whose sum is $s-r+2$. There are $\binom{s-r+1}{h}$ ways to do this. Replace a_0 and a_h by one less, so that the numbers now add up to $s-r$, with a_0 and a_m possibly being zero. Next, arrange ν_1 ones, ν_2 twos, $\dots,\nu_r$ r's in order, as a sequence $(b_1,\dots,b_r)$. There are $\binom{h}{\nu_1\cdots\nu_r}$ different sequences of this sort. Finally, get the set of r numbers as follows: skip $1,\dots,a_0$; put the next b_1 numbers in the set; skip the next a_1 numbers; put the next b_2 in the set; and so forth. The product in (16) counts all such sets.

Therefore, with z_p defined by (15),

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

The sum,

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

represents the *Bell polynomial* $Y_r(\lambda;z_1,2!z_2,\dots,r!z_r)$. Since $\sum_{k=1}^{\infty}z_kt^k=tv^T\mathbb{A}[\mathbb{I}-t\mathbf{V}\mathbb{A}]^{-1}v$, one gets from formula (45) in [12],

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

The definition of $\mathbf{V}=(\mathbb{U}-\mathbb{I})\mathbb{B}^{-1}$ gives

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

Therefore, since $v=(\mathbb{U}-\mathbb{I})\mathbb{B}^{-1}e=\mathbb{B}^{-1}(\mathbb{U}-\mathbb{I})e$, we have

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

and

$$M(\mathbb{U}) = \exp\left\{\lambda v^T e + \lambda v^T \mathbb{A}[\mathbb{I} - \mathbf{V}\mathbb{A}]^{-1} v\right\}$$

$$= \exp\left\{\lambda[e^T \mathbb{B}^{-1}(\Pi - \mathbb{I})e + e^T \mathbb{B}^{-1}(\mathbb{I} - \Pi)(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e]\right\}. \tag{17}$$

Condition (12) now shows that

$$\log M(\mathbb{U}) = \lambda[e^T \mathbb{B}^{-1}(\mathbb{I} - \Pi)(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}(\mathbb{I} - \Pi)e - 1].$$

The matrix Π has nonnegative elements, and all coordinates of the vectors $d = (I - \Pi)e$ and $c = (\mathbb{I} - \Pi^T)\mathbb{B}^{-1}e$ are nonnegative. Thus, (13) corresponds to a compound Poisson distribution, i.e., $\log M(\mathbb{U}) = \lambda[\Phi(\mathbb{U}) - 1]$, where $\Phi(\mathbb{U}) = c^T(\mathbb{I} - \mathbb{U}\Pi)^{-1}\mathbb{U}d = \sum_{k=0}^{\infty} c^T(\mathbb{U}\Pi)^k \mathbb{U}d$ is the probability generating function of the compounding probability distribution.

4. Discussion. When $\Omega = \{\imath\}$, Theorem 1 has been proven in [3] for independent and identically distributed variables. The condition there, $\mathbb{C}_{\imath\imath}(1) \to \mu = 1/\beta_\imath > 1$, means that $\Pi = 1 - \beta_\imath$, and (10) holds automatically. See also [19] for a related result, which is employed in the overlapping template matching test of randomness in [15]. The existence of the limit in (9) is a classical condition on *rare* events whose probabilities are of the order $O(n^{-1})$ or whose length m is of the order $\log n$.

Condition (11) means that with $\mathbb{C}_m(1)$ denoting the restriction of $\mathbb{C}(1)$ in (2) to Ω_m,

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

The probability generating function of the compounding discrete vector, which determines the multivariate Pòlya–Aeppli law in (13), can be written as

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

For this distribution the mean vector is $\mathbb{B}^{-1}e$, and the covariance matrix is

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

When Ω is a one-element set, the compounding distribution is merely the geometric law on positive integers, i.e.,

$$\Phi(u) = \frac{u(1-\pi)}{1 - u\pi}, \tag{18}$$

so that $M(u)$ corresponds to the classical Pòlya–Aeppli distribution.

As in the one-dimensional case, the multivariate Pòlya–Aeppli distribution can be represented as that of the sum, $\sum_1^M Y_j$, where the random variable M has a Poisson distribution with parameter λ and is independent of independent identically distributed discrete random vectors Y_j whose common distribution is the compounding law. In other words, the number M of clusters or clumps has a Poisson distribution, while their multidimensional "sizes" follow the compounding distribution.

It follows that $\mathbf{E}\,\omega_\imath = \lambda/\beta_\imath$ and the covariance matrix of pattern frequencies is $\lambda\mathbb{G}$.

An equivalent representation of the p-variate Pòlya–Aeppli distribution is

$$\sum_{\mathbf{k}\geqq\mathbf{0}} P_{\mathbf{k}}\mathbf{k}, \qquad \mathbf{k}^T = (k_1, \ldots, k_p), \tag{19}$$

where $P_{\mathbf{k}}$ denote independent Poisson random variables with parameters $\lambda\gamma_{\mathbf{k}}$. Here $\gamma_{\mathbf{k}}$ are the coefficients of the power series expansion of $\Phi(U)$, i.e., the probabilities of the compounding distribution. Clearly, $\gamma_{0,\dots,1,\dots,0} = c_\jmath d_\jmath$. In the general case, these probabilities can be found from the multivariate Lagrange expansion (see [7, section 1.2.9]). If $\mathrm{Adj}(\mathbb{I} - \mathbb{U}\Pi)$ denotes the adjoint matrix to $\mathbb{I} - \mathbb{U}\Pi$, then

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

According to the mentioned Lagrange formula (see also formula (3) in [6]), if the diagonal matrix $\mathbb{W}$ is formed by the entries $x_i/(e_i^T\Pi x)$, then $\gamma_{\mathbf{k}}$ is the coefficient of the power x series expansion of the function,

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

of $x^T = (x_1, \dots, x_p)$ (which is supposed to admit a power series expansion).

For example, if $\Pi = ab^T$ with positive vectors a, b satisfies the conditions of Theorem 1, then

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

as $\mathrm{Adj}(\mathbb{I} - (xb^T)/b^T x) = (xb^T)/b^T x$. It follows that when $\sum k_i \geq 2$,

$$\begin{aligned} \gamma_{\mathbf{k}} = &\sum_\jmath \binom{\sum k_i - 2}{k_1 \cdots k_\jmath - 2 \cdots k_p} (a_1 b_1)^{k_1} \cdots (a_\jmath b_\jmath)^{k_\jmath - 1} \cdots (a_p b_p)^{k_p} c_\jmath d_\jmath \\ &+ \sum_{\jmath < \ell} \binom{\sum k_i - 2}{k_1 \cdots k_\jmath - 1 \cdots k_\ell - 1 \cdots k_p} \\ &\times (a_1 b_1)^{k_1} \cdots (a_\jmath b_\jmath)^{k_\jmath - 1} \cdots (a_\ell b_\ell)^{k_\ell - 1} \cdots (a_p b_p)^{k_p} (a_\jmath c_\jmath b_\ell d_\ell + a_\ell c_\ell b_\jmath d_\jmath). \end{aligned}$$

Parameters λ and $\beta_\imath$ are determined by (9) only up to a common factor. To completely specify these parameters one can invoke the normalization condition (12), which means that $\Phi(0) = 0$, or that the support of the compounding discrete multivariate distribution does not include the origin. This condition can be violated for the marginal distributions of the compounding law, although it is clear from derivation of (17) that this representation is also true for subsets of patterns in Ω_m. If $\mathbb{U}_1$ is the subvector of $\mathbb{U}$, corresponding to the selected patterns, the probability generating function, $\Phi_1(\mathbb{U}_1)$, of the first components of the vector ω is $\Phi(\mathbb{U}_1, \mathbb{I})$. Partition

$$\Pi = \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{pmatrix},$$

with similarly partitioned vectors c and d.

Let $\Pi_1 = \Pi_{11} + \Pi_{12}(\mathbb{I} - \Pi_{22})^{-1}\Pi_{21}$, so that $\Pi_1 = \mathbb{I} - [\mathbb{I} + (\Pi(\mathbb{I} - \Pi)^{-1})_{11}]^{-1}$. Then for appropriate $\widetilde{c}$, $\widetilde{d}$,

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

It follows that any univariate marginal is a mixture of a geometric distribution on positive integers and a point mass at zero,

$$\Phi(u_\jmath) = \frac{u_\jmath(1-\pi_\jmath)}{1-u_\jmath\pi_\jmath} + 1 - c_\jmath = \frac{1-c_\jmath+u_\jmath(c_\jmath-\pi_\jmath)}{1-u_\jmath\pi_\jmath},$$

where $0 \le c_\jmath \le 1$, $0 \le \pi_\jmath \le 1$.

As the sum of compounding distribution components does not have to be a geometric random variable, the asymptotic distribution of the sum of frequencies $S = \sum_{\imath\in\Omega}\omega_\imath$ may not have the one-dimensional Pòlya–Aeppli distribution, while their joint multivariate distribution belongs to this class. However, (18) for S holds with $\pi = b^T a$ if $\Pi = ab^T$ with vectors a, b as above.

Theorem 1 is valid under a weaker version of (12), namely if $c^T e \le 1$. However, then the compounding distribution is given by the generating function,

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

The proof of Theorem 1 shows that the joint distribution of all frequencies $\omega_\imath$, $\imath \in \mathbb{Q}^\infty$, weakly converges to the *Pòlya–Aeppli stochastic process*. This process can be defined by a countably infinite substochastic matrix Π_∞ such that $I - \Pi_\infty$ has an inverse operator in the space of all bounded sequences, with the vector $d = (\mathbb{I}-\Pi_\infty)e$ having nonnegative coordinates, and by a probability vector c. The finite-dimensional distributions corresponding to patterns in Ω_m have the probability generating function of the form (20) with c and d denoting restrictions onto the corresponding subspace, and $\Pi = \mathbb{I} - [\mathbb{I} + \Pi_\infty(I-\Pi_\infty)^{-1}|_{\Omega_m}]^{-1}$.

If $\Pi = w\Xi$, with a transition probabilities matrix Ξ and a scalar w, $0 < w < 1$, then $d = (\mathbb{I}-\Pi)e = (1-w)e$ and

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

where $\Phi_N(\mathbb{U}) = c^T(\mathbb{U}\Pi)^{N-1}\mathbb{U}e = c^T\mathbb{U}(\Pi\mathbb{U})^{N-1}e$ is the probability generating function in (8). The latter corresponds to the count vector for the Markov chain of length N with the transition probabilities matrix Ξ, when the initial distribution is c. In other words, the compounding distribution itself is a mixture of pattern frequency distributions in Markov chains of random, geometrically distributed length.

If Π is a diagonal matrix, then $\log M(\mathbb{U})$ is the sum of $\log M(u_\jmath)$ in (18) with parameters $\lambda_\jmath = \lambda c_\jmath d_\jmath$ and $\pi_\jmath = \Pi_{\jmath\jmath}$, provided that $\sum_\jmath c_\jmath d_\jmath(1-\Pi_{\jmath\jmath})^{-1} = 1$. In particular, if Ω_m is formed by aperiodic, uncorrelated patterns $\imath^m$, such that $\mathbb{C}_{\imath^m\imath^m}(1) = 1$ and $\mathbb{C}_{\imath^m\jmath^m}(1) = 0$, $\imath^m \ne \jmath^m$, then $\Pi = 0$, and the asymptotic distribution of $\omega_\imath, \imath \in \Omega$, is the product of classical Poisson laws. If Π is a block-diagonal matrix, then $\omega_\imath$ and $\omega_\jmath$ for $\imath$ and $\jmath$ from different blocks are independent (and the reverse is true as well).

Upper bounds on the total variation distance for rare pattern counts converging to a compound Poisson distribution were originally obtained in [1]. One may hope for derivation of a stronger version of Theorem 1, which would have a sharp bound on a distance between the multivariate Pòlya–Aeppli distribution and the joint distribution of frequencies possibly by using Stein's method as discussed in [2]. To employ this method, the representation (19) may be useful, but there are technical difficulties. The known results in this direction (e.g., [11], [4], [5], [14]) do not seem to provide estimates leading to the Pòlya–Aeppli law convergence under conditions (9)–(11).

5. Examples. We start this section with uniformly distributed independent $\varepsilon_1, \dots, \varepsilon_n$ when $q = 2$, assuming that as $n \to \infty$, $n/2^m \to \lambda/\beta$, i.e., $\beta_\imath \equiv \beta$. This is a model appropriate for randomness testing. Condition (12) means that $\beta = e^T(\mathbb{I} - \Pi)e$.

For an even $m = 2k$, let $\imath^m = (1, 0, 1, 0, \dots, 1, 0)$, $\jmath^m = (0, 1, 0, 1, \dots, 0, 1)$, and $\ell^m = (1, 1, 1, \dots, 1)$ be three pattern sequences. Then

$$\mathbb{C}_{\imath^m\imath^m}(z) = \sum_{r=0}^{k} \frac{z^{2r}}{2^{2r}} = \mathbb{C}_{\jmath^m\jmath^m}(z), \qquad \mathbb{C}_{\ell^m\ell^m}(z) = \sum_{r=0}^{m} \frac{z^r}{2^r},$$

$$\mathbb{C}_{\imath^m\jmath^m}(z) = \mathbb{C}_{\jmath^m\imath^m}(z) = \sum_{r=1}^{k} \frac{z^{2r-1}}{2^{2r-1}},$$

$$\mathbb{C}_{\imath^m\ell^m}(z) = \mathbb{C}_{\jmath^m\ell^m}(z) = \mathbb{C}_{\ell^m\jmath^m}(z) = \mathbb{C}_{\ell^m\imath^m}(z) = 0.$$

Therefore,

$$\mathbb{C}_m - \mathbb{I} \to \frac{1}{3}\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

and

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

$\beta = \frac{3}{2}$. According to Theorem 1,

$$\mathbf{E}\, u_1^{\omega_\imath} u_2^{\omega_\jmath} u_3^{\omega_\ell} \to \exp\left\{\frac{2\lambda}{3}\left[\frac{u_1 + u_2 + 2u_1u_2 - 4}{4 - u_1u_2} + \frac{u_3 - 1}{2 - u_3}\right]\right\}.$$

Thus, $\imath$, $\jmath$ are asymptotically independent of ℓ (which has the classical Pòlya–Aeppli distribution).

If now $\imath^m = (0, 0, 1, \dots, 1, 1)$ and $\jmath^m = (0, 1, 1, \dots, 1, 1)$, then for the same ℓ as above,

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

and $\beta = \frac{3}{2}$. Thus,

$$\mathbf{E}\, u_1^{\omega_\imath} u_2^{\omega_\jmath} u_3^{\omega_\ell} \to \exp\left\{\frac{\lambda}{6}\left(\frac{u_1u_2u_3 + u_2u_3}{2 - u_3} + 2u_1 + u_2 + u_1u_2 - 6\right)\right\}.$$

If $\kappa_1^p = (0, 1, \dots, 1)$, $\kappa_2^p = (1, 1, \dots, 1, 0)$ are two words of length p, $\imath^{rm} = (\kappa_1, \kappa_1, \dots, \kappa_1)$ and $\jmath^{rm} = (\kappa_2, \kappa_2, \dots, \kappa_2)$ are repetitions of these words r times, so that $m = pr$, then for $\Omega = \{\imath, \jmath\}$,

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

Thus

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

With $\beta = \frac{3}{2} - 2^{-(p-1)}$,

$$\mathbf{E}\, u_1^{\omega_i} u_2^{\omega_j} \to \exp\left\{\lambda\left[\frac{2^{p-1}-1}{\beta(2^p - u_1u_2)}\left(\frac{(2^{p-1}-\beta)u_1u_2}{2^{p-1}-1} + u_1 + u_2\right) - 1\right]\right\}.$$

This example suggests employing Theorem 1 to expand the overlapping template matching test in [15] by considering several patterns (say, $\imath^m = (1,0,1,0,\dots,1,0)$ and ℓ^m) in a new test of randomness.

Consider now a Markov chain for a fixed order, say, one. In DNA applications $q = 4$; let Ω, for example, be a collection of patterns (motif) for which $\Omega_m = \{(i_1,\dots,x,\dots,i_m), x = a,c,g,t\}$ for fixed $i_1,\dots,i_{s-1},i_{s+1},\dots,i_m$ with $p = q$. Assume that the transition probabilities matrix $\Xi = \Xi_m$ is known (well estimated) and that the vector ξ represents both the initial and the stationary distributions, $\Xi^T\xi = \xi$.

A prelimiting version of condition (11) can be verified by evaluating the matrix $\mathbb{C}_m$ via the following formula (cf. (10) in [17]) for the elements of $\mathbb{C}_m(1)$:

$$\mathbb{C}(\imath^m, \jmath^m) = \delta_{\imath^m\jmath^m} + \sum_{r=1}^{m-1} \delta_{(i_{r+1}\dots i_m),(j_1\dots j_{m-r})}\xi_{j_{m-r}j_{m-r+1}}\cdots\xi_{j_{m-1}j_m}$$
$$+ \xi_{j_1j_2}\cdots\xi_{j_{m-1}j_m}[\Xi(\mathbb{I} - \Xi + e\xi^T)^{-1}]_{i_mj_1} - P(\jmath^m).$$

If the matrix $\mathbb{K}_m$ is formed by the entries

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

then

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

Indeed,

$$\xi_{j_1j_2}\cdots\xi_{j_{m-1}j_m}[\Xi(\mathbb{I} - \Xi + e\xi^T)^{-1}]_{i_mj_1} - P(\jmath^m)$$
$$= \xi_{j_1j_2}\cdots\xi_{j_{m-1}j_m}\sum_k \xi_{i_mk}\left[(\mathbb{I} - \Xi + e\xi^T)^{-1}_{kj_1} - \xi_{j_1}\right],$$

so that with $\zeta = \max\xi_{rs} < 1$,

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

Thus, (11) is to be verified for the matrix $\mathbb{I}+\mathbb{K}_m$. If this condition is met and the length n of the underlying sequence is such that the product $nP(\imath) = n\xi_{i_1}\xi_{i_1i_2}\cdots\xi_{i_{m-1}i_m}$ is sufficiently large, one can put with $p = \{P(\imath)\}$,

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

Then the multivariate Pòlya–Aeppli law can be employed to find regions where the patterns occur too often or are too rare.

A straightforward calculation in the example of section 7.1 of [13] shows that with $m = 8$, $i_1 = i_4 = i_5 = i_7 = i_8 = g, i_3 = i_6 = t, s = 2$, $n = 1,830,140$,

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

One gets $d = 0.9881e$ and with $a = [0, 0, 0.0117, 0.0001]^T$,

$$\Pi = ea^T.$$

The discussion of the previous section gives the probabilities $\gamma_{\mathbf{k}}$ of the compounding distribution whose bulk (98.46%) is concentrated at $\sum k_i = 1$,

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

The expected frequency of the patterns from this motif is the same as under the Markov chain model (M1) in [13], but in addition the covariance matrix Σ of the frequencies can be obtained. The (usual, not pattern) correlation matrix has the form

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

Similar calculations can be performed for much more complicated motifs, e.g., when Ω_m contains correlated patterns.

Acknowledgments. The author is grateful to Dr. J. Lawrence for the present proof of (16). Helpful comments of a referee are also acknowledged.

REFERENCES

[1] A. D. Barbour, L. Chen, and W. L. Loh, *Compound Poisson approximation for nonnegative random variables via Stein's method*, Ann. Probab., 20 (1992), pp. 1843–1866.

[2] A. D. Barbour and O. Chryssaphinou, *Compound Poisson approximation: A user's guide*, Ann. Appl. Probab., 11 (2001), pp. 964–1002.

[3] O. Chryssaphinou and S. Papastavridis, *A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials*, Probab. Theory Related Fields, 79 (1988), pp. 129–143.

[4] T. Erhardsson, *Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth–death chains*, Ann. Appl. Probab., 10 (2000), pp. 573–591.

[5] T. Erhardsson, *Stein's method for Poisson and compound Poisson approximation*, in An Introduction to Stein's Method, A. Barbour and L. H. Y. Chen, eds., Singapore University Press, Singapore, 2005, pp. 61–113.

[6] I. J. Good, *The frequency count of a Markov chain and the transition to continuous time*, Ann. Math. Statist., 32 (1961), pp. 41–48.

[7] I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, John Wiley, New York, 1983.

[8] L. J. Guibas and A. M. Odlyzko, *Long repetitive patterns in random sequences*, Z. Wahrsch. Verw. Gebiete, 53 (1980), pp. 241–262.

[9] L. J. Guibas and A. M. Odlyzko, *Strings overlaps, pattern matching, and nontransitive games*, J. Combin. Theory Ser. A, 30 (1981), pp. 183–208.

[10] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Van Nostrand, New York, 1967.

[11] G. Reinert and S. Schbath, *Large compound Poisson approximations for occurrences of multiple words*, in Statistics in Molecular Biology and Genetics (Seattle, 1997), IMS Lecture Notes Monogr. Ser. 33, Inst. Math. Statist., Hayward, CA, 1999, pp. 257–275.
[12] J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley, New York, 1958.
[13] S. Robin, F. Rodolphe, and S. Schbath, *DNA, Words and Models: Statistics of Exceptional Words*, Cambridge University Press, Cambridge, UK, 2005.
[14] E. Roquain and S. Schbath, *Improved compound Poisson approximation for the number of occurrences of any rare word family in a stationary Markov chain*, Adv. Appl. Probab., 39 (2007), pp. 128–140.
[15] A. L. Rukhin, *Testing randomness: A suite of statistical procedures*, Theory Probab. Appl., 45 (2001), pp. 111–132.
[16] A. L. Rukhin, *Distribution of the number of words with a prescribed frequency and tests of randomness*, Adv. Appl. Probab., 34 (2002), pp. 775–797.
[17] A. L. Rukhin, *Pattern correlation matrices for Markov sequences and tests of randomness*, Theory Probab. Appl., 51 (2007), pp. 663–679.
[18] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley-Interscience, New York, 2001.
[19] A. M. Zubkov and V. G. Mikhailov, *Limit distributions of random variables connected with long duplications in a sequence of independent trials*, Theory Probab. Appl., 19 (1974), pp. 172–179.
[20] A. M. Zubkov and V. G. Mikhailov, *Repetitions of s-tuples in a sequence of independent trials*, Theory Probab. Appl., 24 (1979), pp. 269–282.