Central Limit Theorem Approximations for the Number of Runs in Markov-Dependent Binary Sequences

George C. Mytalas, Michael A. Zazanis*

Department of Statistics, Athens University of Economics and Business, Athens 10434, Greece

Abstract

We consider Markov-dependent binary sequences and study various types of success runs (overlapping, non-overlapping, exact, etc.) by examining additive functionals based on state visits and transitions in an appropriate Markov chain. We establish a multivariate Central Limit Theorem for the number of these types of runs and obtain its covariance matrix by means of the recurrent potential matrix of the Markov chain. Explicit expressions for the covariance matrix are given in the Bernoulli and a simple Markov-dependent case by expressing the recurrent potential matrix in terms of the stationary distribution and the mean transition times in the chain. We also obtain a multivariate Central Limit Theorem for the joint number of non-overlapping runs of various sizes and give its covariance matrix in explicit form for Markov dependent trials.

Key words: Runs, Markov Chains, Potential Matrix, Central Limit Theorem for Runs

1. Introduction

In a sequence of binary trials, $\{\xi_i\}$ with $\xi_i \in \{0, 1\}$, $i = 1, 2, \ldots$, (success=1 or failure=0) a success run of length k is the occurrence of k consecutive successes. We will not discuss here the vast array of applications of the analysis of runs and patterns in Statistics and Applied Probability; for this

*Corresponding author

Email address: zazanis@aueb.gr (Michael A. Zazanis)

we refer the reader to Balakrishnan and Koutras (2002) and Fu and Lou (2003).

Given a realization of n trials there are several different ways of counting the number of success runs of length k depending, among other considerations, on whether overlapping in counting is allowed or not. A success run of length k occurs at position $m \geq k$ of the binary string $\xi_1, \xi_2, \ldots, \xi_n$ if

$$\xi_{m-k+1} = \xi_{m-k+2} = \cdots = \xi_{m-1} = \xi_m = 1. \tag{1}$$

Counting all the positions $m = k, k = 1, \ldots, n$ for which the above condition holds gives the count $M_{n,k}$ of the number of success runs with overlapping allowed. A *non-overlapping success run of length* k occurs at position m if (1) holds and no non-overlapping run of length k has already occurred in positions $m-k+1$,$m-k+2, \ldots, m-1$. The number of non overlapping runs of length k in a string of n trials is denoted by $N_{n,k}$. Runs of exact size k are also of interest, i.e. k consecutive successes flanked on the left and right by failures. An *exact* run of size k occurs at position $m \geq k$ if, in addition to (1), $\xi_{m-k} = 0$ and $\xi_{m+1} = 0$. We denote the number of exact runs of size k in a string of n trials by $J_{n,k}$. Note that, when the values of the ξ_i's are revealed sequentially, an exact run that occurs at position m will be counted when the value of ξ_{m+1} becomes known. Finally, we say that a run of size *greater than or equal to* k occurs at position m if (1) holds and $\xi_{m-k} = 0$. $G_{n,k}$, denotes the number of success runs of size greater than or equal to k in a string of n trials. In a string of consecutive successes of length greater than k only one run of size greater than k is counted according to this definition. See Fu and Lou (2003) for further clarifications regarding these definitions. To illustrate, if $n = 13$ and $k = 2$, in the binary string 0111101110110 we have the following success run counts: $N_{13,2} = 4$, $G_{13,2} = 3$, $M_{13,2} = 6$, and $J_{13,2} = 1$.

In this paper we consider Markov dependent binary trials. We investigate the asymptotic form of the *joint distribution* of $(N_{n,k}, M_{n,k}, G_{n,k}, J_{n,k})$, as $n \to \infty$ and k is fixed, and show that it obeys a multivariate Central Limit Theorem (CLT). For this purpose we consider an appropriate Markov chain and we express the number of success runs of the kind mentioned as additive functionals based on state visits and state transitions for this chain. The multivariate CLT follows then from standard results for such functionals. The covariance matrix of the limiting normal distribution is expressed in terms of the stationary distribution of the Markov chain and its *recurrent*

potential matrix (also known as the fundamental matrix). This allows for the efficient numerical computation of the covariance matrix for quite general types of Markovian dependence of the binary trial sequence. In the special case of simple two-state Markovian dependence (and of course in the case of Bernoulli trials) we take advantage of the connection that exists between the potential matrix of the chain and the mean transition times between the states of the chain. By straightforward, if somewhat involved, calculations we are able to obtain explicit expressions for the potential matrix, and therefore for the covariance matrix itself, in terms of the parameters of the model. The same technique is used to obtain a multivariate CLT for the number of non-overlapping runs of different sizes, $(N_{n,k_1}, \ldots, N_{n,k_\nu})$, in Markov dependent trials and to compute the corresponding covariance matrix. Besides its intrinsic interest and applications in areas including quality control, randomness tests, and reliability, we indicate applications of this result to the analysis of manufacturing systems with random yield.

Many results exist on the exact distribution of these types of runs for Bernoulli trials and, in some cases, also for Markov-dependent trials. Important unified approaches based on Markov chain techniques include Stefanov and Pakes (1997) where exact results are obtained for runs and more general patterns and Fu and Koutras (1994) which gives the distributions of the run statistics $N_{n,k}, M_{n,k}, G_{n,k}, J_{n,k}$, together with the size of the longest run for non-homogeneous Bernoulli trials. There are also approximations based on limit theorems that establish convergence to Poisson or compound Poisson limits in certain asymptotic regimes and which are especially important in applications. For an overview of these approximations see Barbour and Chryssaphinou (2001).

CLT approximations for the number of runs have a long history. Despite their limited accuracy they remain an important theoretical and practical tool in the analysis of runs. Feller (1968) using arguments based on the Central Limit Theorem for renewal processes, gave an normal approximation for the number of non-overlapping success runs in i.i.d. trials. Setting

$$\mu = \frac{1-p^k}{qp^k} \qquad \text{and} \qquad \sigma^2 = \frac{qp^k - (2k+1)\left(qp^k\right)^2 - qp^{3k+1}}{\left(1-p^k\right)^3}, \tag{2}$$

he shows that $(N_{n,k} - n\mu)/\sigma\sqrt{n} \xrightarrow{d} \mathcal{N}(0,1)$. The same approach is essentially used in Fu et al. (2002) in order to obtain the limiting distribution of the number of successes in success runs of length greater than or equal k in a

sequence of Markov dependent binary trials. Fu and Lou (2007) obtain in the same fashion a CLT approximation for the number of non-overlapping occurrences of a simple or compound pattern in i.i.d. multi-type trials.

A different approach towards establishing the asymptotic normality of the number of runs that has been widely used is via the Hoeffding-Robbins Central Limit Theorem for k-dependent random variables. A CLT for $M_{n,k}$ with i.i.d. Bernoulli trials was obtained along these lines by Godbole (1992) expressing $M_{n,k}$ as a sum of stationary $(k-1)$-dependent indicators. Using a similar approach Hirano et al. (1991) gave explicit results establishing that $(M_{n,k} - (n-k+1)\,p^k)/\sigma\sqrt{n} \xrightarrow{d} N(0,1)$ where

$$\sigma^2 = -p^k\left(1-p^k\right) - 2kp^{2k} + \frac{2p^k\left(1-p^k\right)}{q}. \tag{3}$$

Jennen-Steinmetz and Gasser (1986) also use the CLT for k-dependent random variables in order to obtain a multivariate limiting normal distribution for success and failure runs of various lengths with Bernoulli trials that do not necessarily have the same probability of success but which obey certain asymptotic conditions. The same approach is used in Fu and Lou (2007) in order to obtain normal approximations for the number of overlapping occurrences of simple patterns in multi-type i.i.d. trials and in Makri and Psillakis (2011) which obtains a CLT approximation for $J_{n,k}$.

An alternative and far-reaching approach based on exponential families of random variables has been pioneered by Stefanov (1995) for the analysis of the number of occurrences of runs and patterns in binary trials. By essentially constructing an exponential martingale from the transitions of an appropriate Markov chain, and a stopping time corresponding to the completion of the pattern in question, he is able to derive both exact distributional results and CLT approximations for the joint number of success runs of various sizes and to determine the corresponding limiting covariance matrix. We refer the reader also to Stefanov (2000) and Stefanov and Pakes (1997) for further details.

A very rich literature exists on the wider problem of the number of occurrences of patterns in strings of multi-type trials, typically independent or with markovian dependence (see Reinert, Schbath, and Waterman , 2000 for a review). In the study of the joint distribution of pattern frequencies in strings of multi-type trials in Rukhin (2007) the potential matrix of a Markov chain whose states are words of a given length is used explicitly in

order to obtain the asymptotic covariance matrix in the corresponding CLT. The potential matrix is determined in that case via the *pattern correlation matrix* (as opposed to the results of this paper where it is determined using explicit computations for the mean transition times between states). We refer the reader also to Rukhin (2001) for a detailed study of pattern correlation matrices, including their connection to the potential matrix, and to Rukhin (2010) for explicit formulas involving the potential matrix and asymptotic results establishing convergence to the multivariate Pòlya-Aeppli law.

2. A Markov chain for success runs

Throughout the paper we examine trials with two possible outcomes, 0 and 1. In this section we will assume that the sequence of trials $\{\xi_n; n \in \mathbb{N}\}$ is a two-state Markov chain with transition probability matrix

$$P_M = \begin{bmatrix} q_0 & p_0 \\ q & p \end{bmatrix} \tag{4}$$

where $p_0, q_0, p, q > 0$ and $p_0 + q_0 = p + q = 1$. Consider now another Markov chain $\{X_n; n \in \mathbb{N}\}$ with state space $\mathcal{S} = \{0, 1, \ldots, 2k\}$ and transition probability matrix

$$P = \begin{bmatrix} q_0 & p_0 & & & & & \\ q & & p & & & & \\ \vdots & & & \ddots & & & \\ q & & & & p & & \\ \vdots & & & & & \ddots & \\ q & & & & & & p \\ q & & & & p & & \end{bmatrix}. \tag{5}$$

In the above matrix all elements are zero except for $P_{i0} = q$ for $i = 1, 2, \ldots, 2k$, while $P_{00} = q_0$, $P_{01} = p_0$, $P_{i,i+1} = p$ for $i = 1, 2, \ldots, 2k-1$, and $P_{2k,k+1} = p$. It is easy to see that $\{X_n\}$ is irreducible, aperiodic, and positive recurrent. Its stationary distribution is given by

$$\begin{aligned} \pi_0 &= \frac{q}{q + p_0}, \\ \pi_i &= \frac{p_0 q}{q + p_0} p^{i-1} \quad \text{for } i = 1, \ldots, k, \\ \pi_i &= \frac{p_0 q}{q + p_0} \frac{p^{i-1}}{1 - p^k} \quad \text{for } i = k+1, \ldots, 2k. \end{aligned} \tag{6}$$

Suppose that $P(X_0 = 0) = 1$. The total number of runs in n trials for each of the four different types of runs discussed above can be described by counting the number of visits in various states, or the number of state transitions, of the Markov chain $\{X_n; n \in \mathbb{N}\}$. In fact

$$N_{n,k} = \sum_{m=0}^{n-1} \mathbf{1}(X_m \in \{k, 2k\}) \quad \text{non-overlapping runs in } [0, n-1] \tag{7}$$

$$M_{n,k} = \sum_{m=0}^{n-1} \mathbf{1}(X_m \geq k) \quad \text{overlapping runs in } [0, n-1] \tag{8}$$

$$G_{n,k} = \sum_{m=0}^{n-1} \mathbf{1}(X_m = k) \quad \text{runs of size } \geq k \text{ in } [0, n-1] \tag{9}$$

$$J_{n,k} = \sum_{m=0}^{n-1} \mathbf{1}(X_m = k, X_{m+1} = 0) \quad \text{exact runs in } [0, n-1] \tag{10}$$

Note that the case of Bernoulli trials is obtained within the above framework by setting $p_0 = p$ and $q_0 = q$ in the transition matrix P in (5). We will discuss the Bernoulli case at several places in this article, both because of its simplicity and because of the great interest it presents.

3. Potential matrices and the central limit theorem for countable state-space Markov chains

In this section we state for the sake of completeness some standard results that establish the connection between the recurrent potential matrix for positive recurrent Markov chains with countable state space and the variance constant in the Central Limit Theorem for additive functionals of the sample paths of such chains. Suppose that $\{X_n; n \in \mathbb{N}\}$ is a Markov chain with countable state space $\mathcal{S}$ and transition probability matrix P assumed to be irreducible and positive recurrent. Denote by π the corresponding invariant distribution. Let, as usual, $P_{ij}^n := P(X_n = j | X_0 = i)$ and define the *recurrent potential matrix, Z*, (also known as the fundamental matrix) via

$$Z_{ij} = \sum_{n=1}^{\infty} \left(P_{ij}^n - \pi_j \right) + \delta_{ij}, \qquad i, j \in \mathcal{S} \tag{11}$$

where δ_{ij} is the Kroneker symbol which is equal to 1 if $i = j$ and 0 otherwise. The convergence of the series in (11) is a standard result (see for instance Brémaud , 1997). If $f : \mathcal{S} \to \mathbb{R}$ is such that $\mu := \sum_{i\in\mathcal{S}} \pi_i f(i) < \infty$ then the a.s. convergence of $\frac{1}{n}\sum_{m=0}^{n-1} f(X_m)$ to μ is a consequence of the Strong Law of Large Numbers for Markov chains with countable state space (see Brémaud , 1997). The corresponding Central Limit Theorem is given in the following

Theorem 1. *With the above assumptions on the Markov chain* $\{X_n\}$*, let* $f = (f_1, \ldots, f_d)$ *be a function* $\mathcal{S} \to \mathbb{R}^d$ *such that* $\sum_{i\in\mathcal{S}} \pi_i f_k(i) = \mu_k < \infty$ *and* $\sum_{i\in\mathcal{S}} \pi_i f_k^2(i) < \infty$ *for* $k = 1, 2, \ldots, d$*, and define the additive functional* $\{S_n = (S_{n,1}, \ldots, S_{n,d}); n \in \mathbb{N}\}$ *via* $S_n = \sum_{m=0}^{n-1} f(X_m)$*. Then, with* $\mu = (\mu_1, \ldots, \mu_d)$*, we have*

$$n^{-1/2}\,(S_n - n\mu) \xrightarrow{d} N(0, V)$$

where the covariance matrix is given by

$$V_{kl} = \sum_{(i,j)\in\mathcal{S}\times\mathcal{S}} f_k(i)\Gamma_{ij} f_l(j), \qquad k, l = 1, \ldots, d, \tag{12}$$

with

$$\Gamma_{ij} = \pi_i Z_{ij} + \pi_j Z_{ji} - \pi_i\pi_j - \delta_{ij}\pi_i, \qquad i, j \in \mathcal{S}. \tag{13}$$

For a proof we refer the reader to Aldous and Fill (1997, Ch.2) and in a form where the appearance of the potential matrix is implicit in Port (1994, p.823). When the state space $\mathcal{S}$ is finite, say consisting of n elements, the potential matrix is given by the expression

$$Z = (I - P + \Pi)^{-1} \tag{14}$$

where

$$\Pi := \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \vdots & & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{bmatrix}$$

is a matrix with n identical rows, each equal to the stationary distribution π, and I is the identity matrix. Equation (14) provides an efficient way for the computation of the recurrent potential matrix by means of numerical methods.

An interesting connection exists between the elements of the potential matrix and the mean transition times between states of the chain. If $T_i := \inf\{n > 0; X_n = i\}$ then the following relations hold.

$$Z_{ii} = \pi_i E_\pi T_i, \tag{15}$$
$$Z_{ij} = Z_{jj} - \pi_j E_i T_j. \tag{16}$$

(For a proof we refer the reader to Brémaud , 1997). In (15), (16), $E_i T_j$ denotes the expectation $E[T_j | X_0 = i]$ whereas $E_\pi T_i$ denotes the expectation of T_i given that X_0 is distributed according to the stationary distribution π. When the transition probability matrix has special structure one may exploit (15) and (16) in order to obtain the elements of the potential matrix in closed form. In this paper we will follow this route for Markov dependent trials with dependence given by (4).

3.1. The transition chain

We now define a new Markov chain with state space $\mathcal{S} \times \mathcal{S}$, called the *transition chain*, by setting $Y_n := (X_n, X_{n+1})$ for all n. The fact that $\{Y_n; n \in \mathbb{N}\}$ is a Markov process is immediate. The transition probability matrix of the new chain is given, in terms of the old one, by $P_{(i_1,j_1),(i_2,j_2)} = \delta_{j_1 i_2} P_{i_2 j_2}$. Furthermore, the process $\{Y_n\}$ inherits the properties of irreducibility and positive recurrence from $\{X_n\}$. The stationary distribution of the new chain, $\{\pi(i,j);\ (i,j) \in \mathcal{S} \times \mathcal{S}\}$, is given in terms of the transition probability matrix and the stationary distribution of the original chain by

$$\pi(i,j) = \pi_i P_{ij}. \tag{17}$$

Proposition 2. *The potential matrix of the transition chain can be obtained in terms of that of the original chain as follows:*

$$Z_{(ij)(kl)} = \delta_{ik}\delta_{jl} - \pi_k P_{kl} + Z_{jk} P_{kl}. \tag{18}$$

Proof: The proposition follows by a straightforward computation which we sketch for the sake of completeness. It is simply a matter of evaluating the infinite series

$$Z_{(ij)(kl)} = \sum_{n=1}^{\infty} \left(P^n_{(ij)(kl)} - \pi(k,l) \right) + \delta_{(i,j)(k,l)}.$$

Note that

$$P^n_{(ij)(kl)} = P^{n-1}_{jk} P_{kl} \qquad \text{for } n = 1, 2, 3, \ldots,$$

where, of course, $P^0_{jk} = \delta_{jk}$. Hence

$$Z_{(ij)(kl)} = \left(\delta_{jk} - \pi_k + \sum_{n=1}^{\infty} \left(P^n_{jk} - \pi_k \right) \right) P_{kl} + \delta_{ik}\delta_{jl}.$$

From the above (18) follows readily. ∎

Suppose now that $g : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$ is a reward function based on transitions and consider the additive functional

$$T_n = \sum_{m=0}^{n-1} g(X_m, X_{m+1}).$$

If $\nu := \sum_{(i,j)\in\mathcal{S}\times\mathcal{S}} \pi_i P_{ij} g(i,j)$ exists then we have the Strong Law of Large Numbers $\frac{1}{n} T_n \to \nu$ w.p. 1. The corresponding CLT for the transition chain is given in the following

Theorem 3. *With the above assumptions on the Markov chain $\{X_n\}$, suppose that $\sum_{(k,l)\in\mathcal{S}\times\mathcal{S}} \pi_k P_{kl} g_i^2(k,l) < \infty$ for $i = 1, \ldots, d$. Then*

$$n^{-1/2}(T_n - n\nu) \xrightarrow{d} N(0, \Upsilon),$$

where

$$\Upsilon_{ij} = \sum_{(l_1,l_2),(k_1,k_2)\in\mathcal{S}\times\mathcal{S}} g_i(l_1,l_2) \Gamma_{(l_1,l_2),(k_1,k_2)} g_j(k_1,k_2), \quad i,j = 1,\ldots,d. \tag{19}$$

and

$$\begin{aligned} \Gamma_{(l_1,l_2),(k_1,k_2)} \;=\; & \pi_{(l_1,l_2)} Z_{(l_1,l_2),(k_1,k_2)} + \pi_{(k_1,k_2)} Z_{(k_1,k_2),(l_1,l_2)} \\ & - \pi_{(l_1,l_2)} \pi_{(k_1,k_2)} - \delta_{(l_1,l_2),(k_1,k_2)} \pi_{(l_1,l_2)}. \end{aligned} \tag{20}$$

This is of course a direct restatement of theorem 1 for the transition chain and its proof is omitted.

If we use the expressions for $\pi_{(l_1,l_2)}$ and $Z_{(l_1,l_2),(k_1,k_2)}$ in (17) and (18) we can express the matrix Γ of the transition chain in terms of the transition

probabilities and stationary distribution of the original Markov chain as follows

$$\begin{aligned}\Gamma_{(l_1,l_2),(k_1,k_2)} &= \pi_{l_1}P_{l_1l_2}\left(P_{k_1k_2}Z_{l_2k_1}-\pi_{k_1}P_{k_1k_2}\right)+\pi_{k_1}P_{k_1k_2}\left(P_{l_1l_2}Z_{k_2l_1}-\pi_{l_1}P_{l_1l_2}\right)\\ &\quad -\pi_{l_1}\pi_{k_1}P_{l_1l_2}P_{k_1k_2}, \qquad \text{for } (l_1,l_2)\neq(k_1,k_2) \qquad (21)\\ \Gamma_{(l_1,l_2),(l_1,l_2)} &= 2\pi_{l_1}P_{l_1l_2}\left(1-\pi_{l_1}P_{l_1l_2}+P_{l_1l_2}Z_{l_2l_1}\right)-\left(\pi_{l_1}P_{l_1l_2}\right)^2-\pi_{l_1}P_{l_1l_2}. \qquad (22)\end{aligned}$$

We close this section with the consideration of the joint asymptotic distribution of two additive functionals, one based on state visits and the other based on state transitions. Let $f:\mathcal{S}\to\mathbb{R}$ and $g:\mathcal{S}\times\mathcal{S}\to\mathbb{R}$ and consider two such S_n^f, and S_n^g defined by

$$S_n^f := \sum_{m=0}^{n-1} f(X_m), \qquad S_n^g := \sum_{m=0}^{n-1} g(X_m, X_{m+1}). \tag{23}$$

Clearly, by defining the function $\tilde{f}:\mathcal{S}\times\mathcal{S}\to\mathbb{R}$ via $\tilde{f}(i,j)=f(i)$ for all $(i,j)\in\mathcal{S}\times\mathcal{S}$ the above case is covered by theorem 3. Nevertheless, from a computational point of view, it is worthwhile to examine this case separately, from first principles. The above argument shows that a CLT holds for (S_n^f, S_n^g) and the asymptotic variances of S_n^f and S_n^g can be obtained from theorems 2 and 3 respectively, so the only remaining issue is the determination of the asymptotic covariance. This is given in the following

Proposition 4. *Let S_n^f, S_n^g be defined as in (23) and suppose that $\sum_{i\in\mathcal{S}} f^2(i)\pi_i<\infty$ and $\sum_{i,j\in\mathcal{S}} g^2(i,j)\pi(i,j)<\infty$. Then*

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(S_n^f,S_n^g)=\sum_{i,j,k\in\mathcal{S}} g(i,j)f(k)\left(\pi_iP_{ij}Z_{jk}+\pi_kZ_{ki}P_{ij}-2\pi_iP_{ij}\pi_k\right). \tag{24}$$

Proof: Begin by expressing the asymptotic covariance as

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(S_n^f,S_n^g)=\lim_{n\to\infty}\frac{1}{n}E_\pi(S_n^fS_n^g)-(E_\pi S_n^f)(E_\pi S_n^g),$$

where E_π denotes expectation with respect to the stationary distribution. From (23) we can see that

$$\begin{aligned}E_\pi(S_n^fS_n^g) &= \sum_{m=0}^{n-1}\sum_{l=m+1}^{n-1}\sum_{i,j,k\in\mathcal{S}} g(i,j)f(k)\pi_iP_{ij}P_{jk}^{l-m-1}\\ &\quad +\sum_{m=0}^{n-1}\sum_{l=0}^{m}\sum_{k,i,j\in\mathcal{S}} f(k)g(i,j)\pi_kP_{ij}P_{ki}^{l-m}.\end{aligned}$$

Also, $(E_\pi S_n^f)(E_\pi S_n^g) = n^2 \sum_{i,j,k\in\mathcal{S}} g(i,j)f(k)\pi_k\pi_{i,j}$. Thus, subtracting, we have

$$\begin{aligned}\mathsf{Cov}(S_n^f, S_n^g) &= \sum_{i,j,k\in\mathcal{S}} \sum_{m=0}^{n-1} \sum_{l=m+1}^{n-1} g(i,j)f(k)\left(\pi_i P_{ij} P_{jk}^{l-m-1} - \pi_i P_{ij}\pi_k\right) \\ &\quad + \sum_{k,i,j\in\mathcal{S}} \sum_{m=0}^{n-1} \sum_{l=0}^{m} f(k)g(i,j)\left(\pi_k P_{ij} P_{ki}^{l-m} - \pi_k\pi_i P_{ij}\right).\end{aligned}$$

Now the following limits can be easily evaluated in terms of the fundamental matrix Z in view of its definition in (11).

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} \sum_{l=m+1}^{n-1} \pi_i P_{ij}\left(P_{jk}^{l-m-1} - \pi_k\right) = \pi_i P_{ij} Z_{jk} - \pi_i P_{ij}\pi_k, \tag{25}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} \sum_{l=0}^{m} \pi_k P_{ij}\left(P_{ki}^{l-m} - \pi_i\right) = \pi_k Z_{ki} P_{ij} - \pi_i P_{ij}\pi_k. \tag{26}$$

Taking these into account we obtain (24). ∎

4. The potential matrix for Bernoulli trials

In this section, for the sake of simplicity of exposition and due to its intrinsic interest, we will present the analysis for Bernoulli trials. The corresponding results for Markov dependent trials with dependence given by (4) are sketched in the appendix. The transition probability matrix for the process examined in this section is the matrix (5) modified in its first row by setting $p_0 = p$, $q_0 = q$. Conditioning on the first transition one easily obtains the mean transition times $m_{ij} := E_i T_j$ and $E_\pi T_j = \sum_{i=0}^{2k} \pi_i m_{ij}$. Using (15) and (16) we obtain the following expressions for the elements of the potential matrix.

$$Z_{jj} = \begin{cases} 1 - jqp^j, & \text{if } 0 \le j \le k, \\ \dfrac{1}{1-p^k} - \dfrac{jqp^j}{1-p^k} - \dfrac{kqp^{k+j}}{(1-p^k)^2}, & \text{if } j > k. \end{cases} \tag{27}$$

$$Z_{ij} = \begin{cases} -jqp^j + p^{j-i}, & \text{if } i < j,\ \ j \le k, \\ -jqp^j, & \text{if } i > j,\ j \le k, \\ \dfrac{p^{j-i} - jqp^j}{1-p^k} - \dfrac{kqp^{k+j}}{(1-p^k)^2}, & i < j,\ j > k, \\ \dfrac{p^{k+j-i} - jqp^j}{1-p^k} - \dfrac{kqp^{k+j}}{(1-p^k)^2}, & \text{if } i > j,\ j > k. \end{cases} \tag{28}$$

5. Explicit expression for the covariance matrix in terms of the stationary distribution and the potential matrix

In this section we evaluate the elements of the covariance matrix in the multivariate CLT for the various types of runs based on the ideas of section 3 and the results of section 4. The analysis is presented here again for Bernoulli trials and the corresponding results for two-state Markov dependent trials are relegated to the appendix. In subsection 5.1 we give results for the asymptotic variances and covariances of $M_{n,k}$, $G_{n,k}$, and $N_{n,k}$ whereas in subsection 5.2 we present the results for the asymptotic variance of $J_{n,k}$ and its asymptotic covariance with the other types of runs.

5.1. Runs expressed in terms of state visits

Let us use the CLT of Theorem 1 for additive functionals of the form $S_n^f = \sum_{m=0}^{n-1} f(X_m)$ with $f(x) = 1_A(x)$ where A is an appropriate set of states. The asymptotic variance, as given by (12), taking into account (13) becomes

$$\lim_{n\to\infty} \frac{1}{n} \mathsf{Var}(S_n) = 2 \sum_{i,j\in A} \pi_i Z_{ij} - \sum_{i\in A} \pi_i - \sum_{i,j\in A} \pi_i \pi_j. \tag{29}$$

Equations (7), (8), and (9) show that, by choosing appropriately the set A we can obtain the asymptotic variances of the number of success runs $N_{n,k}$, $M_{n,k}$, and $G_{n,k}$.

Asymptotic Variance of $N_{n,k}$. Taking (29) with $A = \{k, 2k\}$ we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(N_{n,k}) &= 2\left(\pi_k Z_{k,k} + \pi_{2k} Z_{2k,2k}\right) + 2\left(\pi_k Z_{k,2k} + \pi_{2k} Z_{2k,k}\right) \\ &\quad -\pi_k(\pi_k+1) - \pi_{2k}(\pi_{2k}+1) - 2\pi_k\pi_{2k} \\ &= \frac{qp^k}{(1-p^k)^3}\left(1-(2k+1)qp^k - p^{2k+1}\right)\end{aligned}$$

which agrees with (2).

Asymptotic Variance of $M_{n,k}$. Taking (29) with $A = \{k, k+1, \ldots, 2k\}$ we obtain

$$\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(M_{n,k}) = \frac{1+p}{q}p^k(1-p^k) - 2kp^{2k}.$$

Indeed, this agrees with (3) given in Hirano et al. (1991).

Asymptotic Variance $G_{n,k}$. Taking (29) with $A = \{k\}$ we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(G_{n,k}) &= 2\pi_k Z_{kk} - \pi_k(\pi_k+1) \\ &= qp^k\left(1-(1+2k)qp^k\right).\end{aligned}$$

When $k = 1$ then the random variable $G_{n,1}$ denotes the total number of success runs of length 1 in sample of size n. In this case we have

$$\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(G_{n,1}) = qp(1-3qp)$$

which is the expression obtained in Theorem 4 in (Fu and Lou , 2007, pg. 201).

We next obtain the asymptotic covariances. From Theorem 1 the asymptotic covariance of $S_n^f = \sum_{m=0}^{n-1} f(X_m)$ and $S_n^g = \sum_{m=0}^{n-1} g(X_m)$ when $f(x) = 1_A(x)$ and $g(x) = 1_B(x)$ is given by

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(S_n^f, S_n^g) &= \sum_{i\in A,\, j\in B}\left(\pi_i Z_{ij} + \pi_j Z_{ji}\right) - \sum_{i\in A\cap B}\pi_i \\ &\quad - \left(\sum_{i\in A}\pi_i\right)\left(\sum_{j\in B}\pi_j\right), \qquad (30)\end{aligned}$$

where we have again taken into account (13).

Again, by appropriate choice of the sets A and B we can obtain the asymptotic covariances for the number of these types of runs.

Asymptotic Covariance of $(M_{n,k}, N_{n,k})$. Using (30) with $A = \{k, 2k\}$ and $B = \{k, k+1, \ldots, 2k\}$ we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(M_{n,k}, N_{n,k}) &= -\left(1 + \sum_{i=k}^{2k} \pi_i\right)(\pi_k + \pi_{2k}) \\ &\quad + \sum_{i=k}^{2k} (\pi_i Z_{ik} + \pi_k Z_{ki} + \pi_i Z_{i,2k} + \pi_{2k} Z_{2k,i}) \\ &= p^k - \frac{kqp^{2k}}{1-p^k}.\end{aligned}$$

Asymptotic Covariance of $(G_{n,k}, M_{n,k})$. Using (30) with $A = \{k\}$ and $B = \{k, k+1, \ldots, 2k\}$ we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(G_{n,k}, M_{n,k}) &= -\sum_{i=k+1}^{2k} \pi_i\pi_k - \pi_k(\pi_k + 1) + 2\pi_k Z_{kk} \\ &\quad + \sum_{i=k+1}^{2k} (\pi_i Z_{ik} + \pi_k Z_{ki}) \\ &= p^k(1-p^k) - 2kqp^{2k}.\end{aligned}$$

Asymptotic Covariance of $(N_{n,k}, G_{n,k})$. Using (30) with $A = \{k, 2k\}$ and $B = \{k\}$ we obtain

$$\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(N_{n,k}, G_{n,k}) = \frac{qp^k}{1-p^k}\left(1 - (k+1)qp^k - \frac{kqp^k}{1-p^k}\right).$$

5.2. Runs expressed in terms of state transitions

Here we apply Theorem 3 on the functional $S_n^g := \sum_{m=0}^{n-1} g(X_m, X_{m+1})$ with $g : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ an indicator function, namely $g(x, y) = \mathbf{1}_A(x, y)$ where $A \subset \mathcal{S} \times \mathcal{S}$. Using (20) we have

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(S_n^g) &= \sum_{(l_1,l_2),(k_1,k_2)\in\mathcal{S}\times\mathcal{S}} \mathbf{1}_A(l_1, l_2)\Gamma_{(l_1,l_2),(k_1,k_2)}\mathbf{1}_A(k_1, k_2) \\ &= 2\sum_{(l_1,l_2),(k_1,k_2)\in A} \pi_{(l_1,l_2)} Z_{(l_1,l_2),(k_1,k_2)} - \left(\sum_{(l_1,l_2)\in A} \pi_{(l_1,l_2)}\right)^2 - \sum_{(l_1,l_2)\in A} \pi_{(l_1,l_2)}.\end{aligned}$$

We now express the quantities referring to the transition chain in terms of the stationary distribution and the potential matrix of the original chain using (18) to obtain

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(S_n^g) &= 2\sum_{(l_1,l_2),(k_1,k_2)\in A}\pi_{l_1}P_{l_1l_2}Z_{l_2k_1}P_{k_1k_2}-3\left(\sum_{(l_1,l_2)\in A}\pi_{l_1}P_{l_1l_2}\right)^2\\ &\quad+\sum_{(l_1,l_2)\in A}\pi_{l_1}P_{l_1l_2}.\end{aligned}$$

The above applies immediately to $J_{n,k}$, which denotes the number of runs of exact length k, in a set of length n, with $A=\{(k,0)\}$. Hence we have

Asymptotic Variance $J_{n,k}$.

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(J_{n,k}) &= 2\pi_kP_{k0}^2Z_{0k}-3\pi_k^2P_{k0}^2+\pi_kP_{k0}\\ &= q^2p^k+(2p-q(2k+1))q^3p^{2k}.\end{aligned}$$

This agrees with the results of (Makri and Psillakis , 2011, Theorem 2.3).

In order to obtain the asymptotic covariance of $J_{n,k}$ with the other types of runs we will use proposition 4 with $f(x)=1_A(x)$, $g(x,y)=1_B(x,y)$, where A and B are appropriate sets of states and transitions respectively $A\subset\mathcal{S}$, $B\subset\mathcal{S}\times\mathcal{S}$. Then

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(S_n^f,S_n^g) &= -2\left(\sum_{(i,j)\in B}\pi_iP_{ij}\right)\left(\sum_{k\in A}\pi_k\right)\\ &\quad+\sum_{(i,j)\in B,k\in A}\left(\pi_iP_{ij}Z_{jk}+\pi_kZ_{ki}P_{ij}\right)\ .\end{aligned}\tag{31}$$

Asymptotic Covariance $(G_{n,k},J_{n,k})$. Using (31) with $A=\{k\}$ and $B=\{(k,0)\}$ we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k},J_{n,k}) &= -2\pi_k^2P_{k0}+\pi_kP_{k0}Z_{0k}+\pi_kZ_{kk}P_{k0}\\ &= q^2p^k\left(1-p^k(2qk+2q-1)\right).\end{aligned}$$

Asymptotic Covariance $(M_{n,k}, J_{n,k})$. Using again (31) for $A = \{k, k+1, \ldots, 2k\}$ and $B = \{(k,0)\}$ we obtain

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(M_{n,k}, J_{n,k}) &= -2\pi_k P_{k0} \sum_{i=k}^{2k} \pi_i + \sum_{i=k}^{2k} (\pi_k P_{k0} Z_{0i} + \pi_i Z_{ik} P_{k0}) \\ &= q^2 p^k \left(1 - p^k(1+2k)\right).\end{aligned}$$

Asymptotic Covariance $(N_{n,k}, J_{n,k})$. Finally (31) with $A = \{k, 2k\}$ and $B = \{(k,0)\}$ gives

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(N_{n,k}, J_{n,k}) &= -2\pi_k P_{k0}(\pi_k + \pi_{2k}) + \pi_k Z_{kk} P_{k0} + \pi_k P_{k0} Z_{0k} \\ &\quad + \pi_{2k} Z_{2k,k} P_{k0} + \pi_k P_{k0} Z_{0,2k} \\ &= \frac{q^2 p^k}{1-p^k}\left(1 - 2qp^k\right) - \frac{kq^3 p^{2k}}{(1-p^k)^2}(2 - p^k).\end{aligned}$$

6. The Central Limit Theorem for the joint number of runs

Define an additive functional of the transitions and the state visits of the Markov chain $\{X_n\}$ via a function $f = (f_1, f_2, f_3, f_4) : (S, S, S, S \times S) \to \mathbb{R}^4$ so that $\sum_{m=0}^{n-1} f(X_m) = (N_{n,k}, M_{n,k}, G_{n,k}, J_{n,k})$. The Strong Law of Large Numbers for Markov chains gives

$$\frac{1}{n}\left(\sum_{m=0}^{n-1} f_1(X_m), \sum_{m=0}^{n-1} f_2(X_m), \sum_{m=0}^{n-1} f_3(X_m), \sum_{m=0}^{n-1} f_4(X_m, X_{m+1})\right) \longrightarrow \mu \quad \text{w.p. 1.}$$

where $\mu = (\mu_1, \mu_2, \mu_3, \mu_4)$ with $\mu_1 = \sum_{i\in\mathcal{S}} \pi_i f_1(i) = \pi_k + \pi_{2k} = \frac{qp^k}{1-p^k}$, $\mu_2 = \sum_{i\in\mathcal{S}} \pi_i f_2(i) = \sum_{i=k}^{2k} \pi_i = p^k$, $\mu_3 = \sum_{i\in\mathcal{S}} \pi_i f_3(i) = \pi_k = qp^k$, $\mu_4 = \sum_{i\in\mathcal{S}} \pi_{i,j} f_4(i,j) = \pi_{k0} = q\pi_k = q^2 p^k$. Then we can summarize the results of this paper for Bernoulli trials in the following

Theorem 5. *With the above definitions the following Central Limit theorem holds for the number of the types of success runs of size k considered.*

$$\frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} \begin{pmatrix} f_1(X_m) \\ f_2(X_m) \\ f_3(X_m) \\ f_4(X_m, X_{m+1}) \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \xrightarrow{d} N(0, V)$$

where V is the variance-covariance matrix whose elements were determined in the previous sections.

A corresponding CLT holds for Markov dependent trials. The elements of the covariance matrix for dependence of the form given in (4) are those given in the Appendix. The corresponding means are given by $\mu_1 = \frac{qp_0p^{k-1}}{(q+p_0)(1-p^k)}$, $\mu_2 = \frac{p_0p^{k-1}}{q+p_0}$, $\mu_3 = \frac{qp_0p^{k-1}}{q+p_0}$, $\mu_4 = \frac{q^2p_0p^{k-1}}{q+p_0}$.

An application to manufacturing processes with random yield.

Consider an unreliable manufacturing cell where an item is produced when the machine successfully completes k consecutive steps, each requiring a unit of time. If a failure occurs at any given step, no item is produced and the process starts again from scratch. Thus the total number of non-overlapping runs, $N_{n,k}$ gives the number of items produced in a period of n time units. Considering this cell as part of a manufacturing process, we are also interested in the flow of items downstream from the cell, to the next stage of the process. If items produced consecutively, with no failures intervening, are batched together then a stream of batches of random size, with random interarrival intervals arrives at the next station. Since the total number of batches in a time period of n units is $G_{n,k}$, knowledge of the asymptotic joint distribution of $(G_{n,k}, N_{n,k})$ can be used to obtain approximations for the production rate and flow times in such processes. Note that, while $\mathsf{Var}(N_{n,k})$ gives the variability of the total number of arrivals at the downstream station in the time period in question, $\mathsf{Cov}(G_{n,k}, N_{n,k})$ and $\mathsf{Var}(G_n n, k)$ can be used to characterize in an aggregate fashion the temporal smoothness of the arrival process which also affects queueing aspects and delays significantly.

7. A general Markov dependent model

Let $\{\xi_n; n \in \mathbb{N}\}$ be a Markov chain with finite state space $\mathcal{S}$ and irreducible transition probability matrix P_B. We partition the state space $\mathcal{S}$ into two disjoint sets, $\mathcal{S}_0$ and $\mathcal{S}_1$, such that $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$ and correspondingly partition P_B into a block matrix

$$P_B \;=\; \begin{bmatrix} Q_0 & P_0 \\ Q_1 & P_1 \end{bmatrix}. \tag{32}$$

Q_0, P_1, are square matrices of dimension $m_0 := |\mathcal{S}_0|$ and $m_1 := |\mathcal{S}_1|$ respectively while P_0 and Q_1 are rectangular $m_0 \times m_1$ and $m_1 \times m_0$ matrices. In order to simplify the analysis we will assume that the substochastic matrix P_1 is *strictly positive,* i.e. that it contains no zero entries.

We now consider the process $\{\mathbf{1}_{\mathcal{S}_1}(\xi_n); n \in \mathbb{N}\}$ which is a sequence of *dependent binary trials* and use the approach of the previous sections in order to obtain approximations for the number of success runs of various types based on the Central Limit Theorem. Suppose we are interested in success runs of size k. Let $\{X_n; n \in \mathbb{N}\}$ be a Markov chain with state space consisting of $m_0 + 2km_1$ elements. It will be convenient to label its states as $\{(i, r); i = 0, 1, 2, \ldots, 2k; r = 1, 2, \ldots, m_0$ if $i = 0$ and $r = 1, 2, \ldots, m_1$ if $i > 0\}$ and to order them lexicographically. The transition probability matrix of $\{X_n\}$ is

$$P = \begin{bmatrix} Q_0 & P_0 & & & & & & \\ Q_1 & & P_1 & & & & & \\ \vdots & & & \ddots & & & & \\ Q_1 & & & & P_1 & & & \\ Q_1 & & & & & P_1 & & \\ \vdots & & & & & & \ddots & \\ Q_1 & & & & & & & P_1 \\ Q_1 & & & & P_1 & & & \end{bmatrix}. \tag{33}$$

Under the assumption of the irreducibility of the matrix P_B and the additional assumption of strict positivity of P_1, the matrix P is again irreducible and aperiodic. The stationary distribution can be obtained by solving the system $\pi = \pi P$ together with the normalization condition. It will be convenient to partition the stationary distribution as $\pi = [\pi_0, \pi_1, \ldots, \pi_k]$ where π_0 is a $1 \times m_0$ row vector and π_i, $i = 1, \ldots, 2k$, are $1 \times m_1$ row vectors. We will denote these row vectors as $\pi_0 := [\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,m_0}]$ and $\pi_i := [\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,m_1}]$, $i = 1, 2, \ldots, 2k$.

If u_0 is a $m_0 \times 1$ *column vector* with all entries equal to 1, and u_1 a $m_1 \times 1$ column vector with all entries equal to 1, then the normalization condition can be written as

$$\pi_0 u_0 + \sum_{i=1}^{2k} \pi_i u_1 = 1. \tag{34}$$

The stationary equations are

$$\begin{aligned}
\pi_0 &= \pi_0 Q_0 + \sum_{i=1}^{k} \pi_i Q_1, && (35)\\
\pi_1 &= \pi_0 P_0, && (36)\\
\pi_i &= \pi_{i-1} P_1, \quad i = 2, 3, \ldots k, \text{ and } i = k+2, \ldots, 2k, && (37)\\
\pi_{k+1} &= \pi_k P_1 + \pi_{2k} P_1. && (38)
\end{aligned}$$

From (36), (37), and (38) we obtain

$$\begin{aligned}
\pi_i &= \pi_0 P_0 P_1^{i-1}, \qquad i = 1, 2, \ldots, k, && (39)\\
\pi_i &= \pi_0 P_0 P_1^{i-1} (I - P_1^k)^{-1}, \qquad i = k+1, k+2, \ldots, 2k, && (40)
\end{aligned}$$

and thus

$$\begin{aligned}
\sum_{i=1}^{2k} \pi_i &= \pi_0 P_0 \left(I + P_1 + \cdots + P_1^{k-1}\right) + \pi_0 P_0 \left(P_1^k + \cdots + P_1^{2k-1}\right) (I - P_1)^{-1}\\
&= \pi_0 P_0 (I - P_1^k)(I - P_1)^{-1} + \pi_0 P_0 P_1^k (I - P_1^k)(I - P_1)^{-1} (I - P_1^k)^{-1}\\
&= \pi_0 P_0 (I - P_1)^{-1},
\end{aligned}$$

where we have used the fact that $(I - P_1)^{-1}$ and $(I - P_1^k)^{-1}$ commute. Hence (35) becomes

$$\pi_0 = \pi_0 \left(Q_0 + P_0 (I - P_1)^{-1} Q_1\right). \tag{41}$$

The normalization condition (34) becomes

$$\pi_0 u_0 + \pi_0 P_0 (I - P_1)^{-1} u_1 = 1. \tag{42}$$

Note that $P_W := Q_0 + P_0 (I - P_1)^{-1} Q_1$ is in fact an $m_0 \times m_0$ stochastic matrix which gives the transition probability matrix of the chain $\{\xi_n\}$ *watched in the set* $\mathcal{S}_0$. (Indeed, if one defines a sequence of stopping times $\{W_n\}$ with $W_0 := \inf\{n \geq 0 : \xi_n \in \mathcal{S}_0\}$ and $W_n := \inf\{n > W_{n-1} : \xi_n \in \mathcal{S}_0\}$ the process $\{Y_n; n = 0, 1, 2, \ldots\}$ defined via $Y_n := \xi_{W_n}$ is a Markov chain with transition probability matrix given by P_W.) The stochastic matrix P_W inherits its irreducibility from that of P_B and thus (41) has an essentially unique positive solution, within a multiplicative constant which can be determined so that (42) is satisfied.

In general, the stationary distribution can be computed numerically from the above equations while the recurrent potential matrix can be computed

numerically using (14) where Π is the $(m_0 + 2km_1) \times (m_0 + 2km_1)$ matrix with constant rows. Its elements, $(\Pi)_{ij}$, $i, j = 1, 2, \ldots, m_0 + m_1 2k$, are given by $(\Pi)_{ij} = \pi_{0,i}$ if $1 \leq i \leq m_0$ and $(\Pi)_{ij} = \pi_{l,i-m_0-lm_1}$ if $m_0 + (l-1)m_1 < i \leq m_0 + lm_1$, $l = 1, 2, \ldots, 2k$.

The number of the various kinds of runs considered can be expressed as additive functionals of the Markov chain $\{X_n\}$ given by the following expressions.

$$\begin{aligned}
N_{n,k} &= \sum_{j=0}^{n-1} \sum_{r=1}^{m_1} \mathbf{1}(X_j = (k, r)) + \mathbf{1}(X_j = (2k, r)) \\
M_{n,k} &= \sum_{j=0}^{n-1} \sum_{i=k}^{2k} \sum_{r=1}^{m_1} \mathbf{1}(X_j = (i, r)) \\
G_{n,k} &= \sum_{j=0}^{n-1} \sum_{r=1}^{m_1} \mathbf{1}(X_j = (k, r)) \\
J_{n,k} &= \sum_{m=0}^{n-1} \sum_{r=1}^{m_1} \sum_{l=1}^{m_0} \mathbf{1}(X_m = (k, r), X_{m+1} = (0, l))
\end{aligned}$$

The multivariate CLT follows then again by Theorems 1 and 3 while the covariance matrix is obtained as in section 5. The sets A and B that appear there are of course more complicated due to the two-dimensional representation of the state space as is clear from the above expressions for the number of runs.

8. Joint asymptotic distribution of non-overlapping runs of different sizes

In this section we illustrate the generality of the method we propose by obtaining the joint asymptotic distribution of non-overlapping success runs of different lengths. This problem has been considered by Jennen-Steinmetz and Gasser (1986) as mentioned in the introduction. Here we obtain analogous results for Markov dependent trials with dependence given by (4).

Let $\{X_n\}$ be a Markov chain with *countable* state space $\mathcal{S} := \{0, 1, 2, \ldots\}$ and transition probability matrix P given by $P_{00} = q_0$, $P_{01} = p_0$, $P_{i,i+1} = p$, $P_{i0} = q$ and all other entries equal to zero. This chain is clearly irreducible,

aperiodic, and positive recurrent with stationary distribution given by

$$\pi_0 = \frac{q}{p_0+q}, \quad \pi_i = \frac{q}{p_0+q} p_0 p^{i-1}, \quad i = 1, 2, \ldots. \tag{43}$$

The corresponding potential matrix can be obtained from (15), (16), by means of elementary calculations similar to those of section 4.

$$Z_{ij} = -\frac{q p_0 p^{j-1}}{p_0+q}\left(\frac{p-p_0}{p_0+q} + j\right) \; + \; \mathbf{1}(i \le j) p^{j-i}. \tag{44}$$

If k_i, $i = 1, 2, \ldots, \nu$, are fixed positive integers with $0 < k_1 < k_2 < \cdots < k_\nu$ we are interested in determining the joint asymptotic distribution of the number of non-overlapping runs $(N_{k_1,n}, N_{k_2,n}, \ldots, N_{k_\nu,n})$, as the number of trials $n \to \infty$. Since

$$N_{k_i,n} = \sum_{m=0}^{\infty} \mathbf{1}(X_m \in A_i) \;\; \text{with } A_i = \{l k_i, l = 1, 2, \ldots\} \tag{45}$$

the joint asymptotic distribution will be multivariate Normal by the CLT for additive functionals of Markov chains in Theorem 1. The asymptotic variances have already been determined in section 5 for Bernoulli trials and in the Appendix in the Markov dependent case. Therefore the only remaining issue is the determination of the asymptotic covariance between, say, $N_{k_1,n}$ and $N_{k_2,n}$. Using (30) and taking into account (45) we have

$$\begin{aligned} \lim_{n\to} \frac{1}{n} \mathsf{Cov}(N_{k_1,n}, N_{k_2,n}) \;\; &= \;\; \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} \pi_{ik_1} Z_{ik_1,jk_2} + \pi_{jk_2} Z_{jk_2,ik_1} \\ &\quad - \;\; \sum_{i=1}^{\infty} \pi_{i\cdot \mathrm{lcm}(k_1,k_2)} \;\; - \sum_{i=1}^{\infty} \pi_{ik_1} \sum_{j=1}^{\infty} \pi_{jk_2}, \end{aligned} \tag{46}$$

where $\mathrm{lcm}(k_1, k_2)$ stands for the least common multiple of k_1 and k_2. The evaluation of the asymptotic covariance using the expression for the stationary distribution (43) and the potential matrix (44) is straight forward. The only term that complicates matters is a term of the form $\sum_{j=1}^{\infty}\sum_{i=1}^{\infty} \mathbf{1}(ik_1 \le jk_2) p^{jk_2} + \mathbf{1}(jk_2 \le ik_1) p^{ik_1}$ which arises due the presence of the indicator function in (44) and which requires separate consideration according to whether k_1 divides k_2, and if not, according to whether the remainder of the integer

division of k_2 by k_1 divides k_1. (We thus have three separate expressions corresponding to these three cases.)

$$\begin{aligned}\lim_{n\to}\frac{1}{n}\mathsf{Cov}(N_{k_1,n},N_{k_2,n}) &= -\left(\frac{qp_0p^{-1}}{q+p_0}\right)^2\frac{p^{k_1+k_2}}{(1-p^{k_1})(1-p^{k_2})}\\ &\times\left(\frac{1+p-p_0}{p_0+q}+\frac{1}{1-p^{k_1}}+\frac{1}{1-p^{k_2}}\right)+\frac{qp_0p^{-1}}{q+p_0}\cdot J \qquad (47)\end{aligned}$$

where

$$J=\begin{cases}\frac{p^{k_2}}{1-p^{k_2}}\left(\frac{k_2}{k_1}\frac{1}{1-p^{k_2}}+\frac{1}{1-p^{k_1}}-1\right) & \text{if } k_2=ak_1\\ \frac{ap^{k_2}}{(1-p^{k_2})^2}+\frac{1}{1-p^{k_1(ac+1)}}\left(\frac{p^{k_1a}-p^{k_1a(c+1)}}{(1-p^{k_1})(1-p^{k_1a})}-p^{k_1(ac+1)}\right) & \text{if } \begin{matrix}k_2=ak_1+b,\\ k_1=bc\end{matrix}\\ \frac{ap^{k_2}}{(1-p^{k_2})^2}+\frac{1}{(1-p^{k_1})(1-p^{k_1a})}\left(p^{k_1a}-\frac{p^{k_1ca}(1-p^{k_1})}{1-p^{k_1(ca+1)}}\right)-\frac{p^{\mathrm{lcm}(k_1,k_2)}}{1-p^{\mathrm{lcm}(k_1,k_2)}} & \text{if } \begin{matrix}k_2=ak_1+b,\\ k_1=bc+d.\end{matrix}\end{cases}$$

A. Explicit expressions for Markov dependent trials

Here we present the results for Markov dependent trials that follow (4).

A.1. The elements of the potential matrix Z_{ij} for Markov dependent trials

$$Z_{jj}=\begin{cases}\frac{1}{p_0p^{-1}}-\frac{q(p-p_0)}{(q+p_0)^2}, & \text{if } j=0,\\ 1+\frac{(1-j)qp_0p^{j-1}}{q+p_0}-\frac{qp_0p^{j-1}}{(q+p_0)^2}, & \text{if } 1\le j\le k.\\ \frac{1}{1-p^k}-\frac{qp_0p^{j-1}}{(q+p_0)\left(1-p^k\right)}\left(j+\frac{p-p_0}{q+p_0}\right)-\frac{kqp_0p^{k+j-1}}{(q+p_0)\left(1-p^k\right)^2}, & \text{if } j>k.\end{cases}$$

$$Z_{ij}=\begin{cases}\frac{p_0p^{j-1}+(1-j)qp_0p^{j-1}}{q+p_0}-\frac{qp_0p^{j-1}}{(q+p_0)^2}, & \text{if } i=0,\ 1\le j\le k,\\ \frac{(q+p_0)p^{j-i}+(1-j)qp_0p^{j-1}}{q+p_0}-\frac{qp_0p^{j-1}}{(q+p_0)^2}, & \text{if } 1\le i<j\le k\\ \frac{q(p_0-p)}{(q+p_0)^2}, & j=0\\ \frac{(1-j)qp_0p^{j-1}}{q+p_0}-\frac{qp_0p^{j-1}}{(q+p_0)^2}, & \text{if } i>j,\ 1\le j\le k.\end{cases}$$

$$Z_{ij} = \begin{cases} \frac{p_0^2 p^{j-1} + (1-j)(q+p_0)qp_0p^{j-1}}{(q+p_0)^2\left(1-p^k\right)} - \frac{kqp_0p^{k+j-1}}{(q+p_0)\left(1-p^k\right)^2}, & \text{if } i=0, \\ \frac{p^{j-i}}{1-p^k} + \frac{(1-j)qp_0p^{j-1}}{(q+p_0)\left(1-p^k\right)} - \frac{qp_0p^{j-1}}{(q+p_0)^2\left(1-p^k\right)} - \frac{kqp_0p^{k+j-1}}{(q+p_0)\left(1-p^k\right)^2} & \text{if } 1 \le i < j, \ \ j \ge k+1, \\ \frac{p^{k+j-i}}{1-p^k} + \frac{(1-j)qp_0p^{j-1}}{(q+p_0)\left(1-p^k\right)} - \frac{qp_0p^{j-1}}{(q+p_0)^2\left(1-p^k\right)} - \frac{kqp_0p^{k+j-1}}{(q+p_0)\left(1-p^k\right)^2} & \text{if } i > j, \ \ j \ge k+1. \end{cases}$$

A.2. The elements of the covariance matrix for Markov dependent trials

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(N_{n,k}) &= \frac{qp_0p^{k-1}}{(p_0+q)(1-p^k)^2}\left(1-p^{k-1}\frac{q^2(q-q_0)-p_0^2}{(p_0+q)^2}-k\frac{2qp_0p^{k-1}}{(p_0+q)(1-p^k)}\right) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(M_{n,k}) &= \frac{p_0p^{k-1}}{p_0+q}\left(\frac{2-q}{q}-\frac{p_0p^{k-1}}{q+p_0}\left(\frac{2p-q}{q}+\frac{2}{q+p_0}\right)\right) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(G_{n,k}) &= \frac{qp_0p^{k-1}}{p_0+q}\left(1-\frac{qp_0p^{k-1}}{(p_0+q)^2}\left(2(p_0+q)k+p+q_0\right)\right) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(J_{n,k}) &= \frac{q^2p_0p^{k-1}}{p_0+q}+\frac{q^3p_0^2p^{2k-2}}{(q+p_0)^2}\left(\frac{2p_0}{q+p_0}-(2k+1)q\right)
\end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(M_{n,k},N_{n,k}) &= \frac{p_0p^{k-1}}{(p_0+q)(1-p^k)}\left(1-\frac{p_0p^{k-1}}{(p_0+q)^2}\left(p_0+q+2(p-p_0)\right)\right. \\
&\qquad \left. +k\frac{qp^{k-1}}{1-p^k}\left(p-\frac{p_0}{p_0+q}(2-p^k)\right)\right) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k},M_{n,k}) &= \frac{p_0p^{k-1}}{p_0+q}\left(1-\frac{qp_0p^{k-1}}{p_0+q}\left(2k+\frac{p}{q}+\frac{p+q_0}{p_0+q}\right)\right) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(N_{n,k},G_{n,k}) &= \frac{qp_0p^{k-1}}{(p_0+q)(1-p^k)}\left(1-\frac{q(p+q_0)p_0p^{k-1}}{(p_0+q)^2}-k\frac{qp_0p^{k-1}(2-p^k)}{(p_0+q)(1-p^k)}\right) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k},J_{n,k}) &= \frac{q^2p_0p^{k-1}}{q+p_0}-\frac{q^2p_0^2p^{2k-2}}{(p_0+q)^2}\left(2kq+\frac{q-p_0}{q+p_0}\right)
\end{aligned}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(M_{n,k},J_{n,k}) = \frac{q^2p_0p^{k-1}}{p_0+q}\left(1-\frac{p_0p^{k-1}}{p_0+q}\left(2k+\frac{p+q_0}{p_0+q}\right)\right)$$

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(N_{n,k},J_{n,k}) = \frac{q^2p_0p^{k-1}}{p_0+q}\left(1-\frac{p_0p^{k-1}}{(p_0+q)(1-p^k)}\left(\frac{2-p^k}{1-p^k}qk+\frac{q-p_0}{q+p_0}\right)\right)$$

References

Aldous, D., Fill, J. Reversible Markov Chains and Random Walks on Graphs. `http://www.stat.berkeley.edu/~aldous/RWG/book.html`

Balakrishnan, N., Koutras, M. V., 2002. Runs and Scans with Applications, Wiley, New York.

Barbour, A.D., Chryssaphinou, O., 2001. Compound Poisson approximation: A user's guide. The Annals of Applied Probability, 11, 3, 964–1002.

Brémaud, P., 1997. Markov Chains, Springer Verlag.

Feller, W. 1968. An Introduction to Probability Theory and Its Applications vol., 3rd ed. John Wiley, New York.

Fu, J.C., Koutras, M.V., 1994. Distribution theory of runs: a Markov chain approach. J. Amer. Statistic. Assoc. 89, 1050-1058.

Fu, J., Lou, W.Y.W., 2003. Distribution Theory of Runs and Patterns and its Applications, World Scientific, Singapore.

Fu, J., Lou, W.Y.W., 2007. On the normal approximation for the distribution of the number of simple or compound patterns in a random sequence of multi-state trials. Methodology and Computing in Applied Probability, 9, 195-205.

Fu, J., Lou, W.Y.W., Bai, Z.D., Li G., 2002. The exact and limiting distributions for the number of successes in success runs within a sequence of Markov-dependent two-state trials. Annals of the Institute of Statistical Mathematics, 54, 719-730.

Godbole, A.P., 1992. The exact and asymptotic distribution of overlapping success runs. Communications in Statistics–Theory and Methods, 21, 4, 953-967.

Hirano, K., Aki, S., Kashiwagi, N., Kuboki, H., 1991. On Ling's binomial and negative binomial distributions of order k. Statistics and Probability Letters, 11, 503-509.

Jennen-Steinmetz, C., Gasser, T., 1986. The asymptotic power of runs tests. Scandinavian Journal of Statistics, 13, 263-269.

Makri, F.S. and Z.M. Psillakis, 2011. On success runs of a fixed length in Bernoulli sequences: Exact and asymptotic results. Computers and Mathematics with Applications, 61, 761-772.

Port, S.C., 1994. Theoretical Probability for Applications. John Wiley.

Reinert, G., Schbath, S. and Waterman, M. 2000. Probabilistic and statistical properties of words: an overview. Journal of Computational Biology, 7, 1-46.

Rukhin, A. 2001. Pattern correlation matrices and their properties. Linear Algebra and its Applications 327, 105-114.

Rukhin, A. 2007. Pattern correlation matrices for Markov sequences and tests for randomness. Theory of Probability and its Applications 51, 663-679.

Rukhin, A. 2010. Joint distribution of pattern frequencies and multivariate Pòlya-Aeppli law. Theory of Probability and its Applications 54, 246-260.

Stefanov, V. T. 1995. Explicit limit results for minimal sufficient statistics and maximum likelihood estimators in some Markov processes: Exponential families approach. Annals of Statistics 23, 1073-1101.

Stefanov, V.T., 2000. On run statistics for binary trials. Journal of Statistical Planning and Inference, 87, 177-185.

Stefanov, V.T., and A.G. Pakes (1997). Explicit distributional results in pattern formation. Annals of Applied Probability 7, (3) 666-678.