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

George C. Mytalas and Michael A. Zazanis *
Department of Statistics
Athens University of Economics and Business
Athens 10434, Greece

Abstract

We consider Markov-dependent multi-type sequences and study various kinds of runs (including overlapping, non-overlapping, exact, etc.) by examining additive functionals based on state visits and transitions in an appropriately constructed Markov chain. We establish multivariate Central Limit Theorems for the number of these runs and obtain the covariance matrix of the limiting multivariate normal distribution in closed form using the potential matrix. Finally we briefly discuss applications of these results in reliability theory and molecular biology.

KEYWORDS: RUNS, MARKOV CHAINS, POTENTIAL MATRIX, MULTI STATE TRIALS, CENTRAL LIMIT THEOREM FOR RUNS.
AMS 2000 SUBJECT CLASSIFICATION: PRIMARY 60E05, SECONDARY 60J10.

1 Introduction

The study of success runs is important both in statistical theory (e.g. hypothesis testing) and in applications of statistics in various areas, most notably quality control, reliability theory, and molecular biology. For a comprehensive review of the literature on runs we refer the reader to Balakrishnan and Koutras [3] and Fu and Lou [10]. In a sequence of binary (success or failure) trials, a success run of length k is the occurrence of k consecutive successes. Given a realization of n trials there are several different ways of counting the number of success runs of length k, depending on whether overlapping in counting is allowed or not. The choice of definition depends on the specific application. The most frequently used enumeration schemes result in the following statistics (citations in parentheses refer to works where these statistics have been defined):

$N_{n,k}$, the number of non-overlapping consecutive k successes until the nth trial (Feller [8]);

$G_{n,k}$, the number of success runs of size greater than or equal to k until the nth trial (Gibbons [15]);

$M_{n,k}$, the number of overlapping consecutive k successes until the nth trial (Ling [20]);

$J_{n,k}$, the number of runs consisting of exactly k successes until the nth trial (Mood [23]).

*Corresponding author. Email: zazanis@aueb.gr

Finally, another run-related statistic frequently of interest is $S_{n,k}$, the number of successes in success runs of length k or greater until the nth trial. In the last decades significant research interest has been focused on the study of the distribution of the number of runs (and more general patterns) in sequences of Markov-dependent multi-type trials.

Let $\xi_0, \xi_1, \ldots, \xi_n$ be trials from a time homogeneous Markov chain on the finite alphabet $\mathcal{S} = \{s_1, s_2, \ldots, s_d\}$. We are interested in the joint statistics of the number of runs of all symbols, of length k_l for s_l, $l = 1, \ldots, d$. These may be counted in any of the ways discussed above, i.e. overlapping, non-overlapping, exact, etc. and give rise to vector counts of dimension d. More specifically, let N^l_{n,k_l} denote the number of non-overlapping runs of s_l of length k_l, and $\boldsymbol{k} := (k_1, \ldots, k_d)$. Then $\boldsymbol{N}_{n,\boldsymbol{k}} = (N^1_{n,k_1}, \ldots, N^d_{n,k_d})$ is the d-dimensional random vector of counts of non-overlapping runs for each symbol in a string of length n. Similarly we define vector counts of runs of the other kinds discussed, namely $\boldsymbol{G}_{n,\boldsymbol{k}}$, $\boldsymbol{M}_{n,\boldsymbol{k}}$, $\boldsymbol{J}_{n,\boldsymbol{k}}$, and $\boldsymbol{S}_{n,\boldsymbol{k}}$, as d-dimensional random vectors. In this paper we present a Central Limit Theorem (CLT) and a corresponding normal approximation for $\boldsymbol{N}_{n,\boldsymbol{k}}$ and $(\boldsymbol{G}_{n,\boldsymbol{k}}, \boldsymbol{M}_{n,\boldsymbol{k}}, \boldsymbol{J}_{n,\boldsymbol{k}}, \boldsymbol{S}_{n,\boldsymbol{k}})$.

Fu [9] studied multi-state trials by examining the joint distribution of runs and patterns using the method of the embedding Markov chain (see [3], [10]). Using the same method Doi and Yamamoto [7] obtained the joint distribution of the number of runs of c symbols in the sequence of trials from an alphabet of $c + 1$ symbols. A recursive method for the evaluation of the joint distribution of $\boldsymbol{N}_{n,\boldsymbol{k}}$, $\boldsymbol{M}_{n,\boldsymbol{k}}$, $\boldsymbol{G}_{n,\boldsymbol{k}}$, and $\boldsymbol{J}_{n,\boldsymbol{k}}$ in a sequence of multi-state trials is given in Han and Aki [16]. They extended the concept of Markov chain embeddable variables of binomial type introduced by Koutras and Alexandrou [18] to Markov chain embeddable variables of multinomial type. Shinde and Kotwal [25] studied the same multivariate distributions together with the multi-type version of $X_{n,k}$, the number of l-overlapping success runs of length k in n trials, by using conditional probability generating functions in the sequence of Markov-dependent multi-type trials. Inoue and Aki [17] develop formulae for the derivation of the probability generating function and the higher order moments of the number of runs of different lengths and different kinds. An approximation based on compound Poisson limit theorems for overlapping runs in multi-type trials is given by Chryssaphinou and Vagelatou [6]. Fu and Lou [11] apply the finite Markov chain embedding technique together with renewal-theoretic techniques to obtain a CLT and a corresponding normal approximation for the number of non-overlapping and overlapping occurrences of a simple or compound pattern in i.i.d. multi-type trials.

Recently Mytalas and Zazanis [24] investigated the joint distribution of $N_{n,k}$, $M_{n,k}$, $G_{n,k}$, and $J_{n,k}$ for Markov dependent binary trials, showed that it obeys a multivariate CLT and obtained a closed form expression for the covariance matrix of the limiting multivariate normal distribution. They also obtained a multivariate CLT for the joint number of non-overlapping runs of various sizes $(N_{n,k_1}, \ldots, N_{n,k_l})$ and its covariance matrix. In this paper we apply the same methodology extending the results to multi-type trials obtaining the asymptotic form of the joint distributions $\boldsymbol{G}_{n,\boldsymbol{k}}$, $\boldsymbol{M}_{n,\boldsymbol{k}}$, $\boldsymbol{J}_{n,\boldsymbol{k}}$, $\boldsymbol{S}_{n,\boldsymbol{k}}$ and $\boldsymbol{N}_{n,\boldsymbol{k}}$. We also provide applications in biological sequence homology and in reliability theory.

2 Potential matrix methodology: Theoretical tools and description

In this section we give a brief discussion of the *potential matrix methodology* for the computation of the covariance matrix in CLT approximations of the number of various types of runs in long strings of Markov dependent trials. This method was introduced in [24], to which we refer the reader for a more detailed exposition.

Let $\{X_n\}$ be a time-homogeneous, irreducible, aperiodic, positive recurrent Markov chain on a finite or countable state space $\mathcal{X}$. Let P denote its transition probability matrix and, as usual, $P_{ij}^n := \mathbb{P}(X_n = j | X_0 = i)$.

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

denote the elements of the recurrent potential matrix of $\{X_n\}$, also known as the fundamental matrix (see [2]). δ_{ij} is the Kroneker symbol, equal to 1 if $i = j$ and 0 otherwise. The convergence of the series under our assumptions is a standard result (see for instance [5]).

Suppose we are interested in obtaining multivariate normal approximations for the joint distribution of the number of ν different kinds of runs (or other related statistics of interest) in a string of n symbols (trials) of a finite alphabet $\{s_1, \ldots, s_d\}$ coming from a markovian source. The proposed methodology involves the following steps:

1. Construct a Markov Chain $\{X_n\}$ on an appropriate (finite or countable) state space and with a properly chosen irreducible, aperiodic, and positive recurrent transition probability matrix so that ν_1 of the run counts can be obtained by counting the number of visits in various states and ν_2 by counting the number of state transitions of $\{X_n\}$ where of course $\nu_1 + \nu_2 = \nu$. This involves the determination of reward function $\boldsymbol{f} = (f_1, \ldots, f_{\nu_1}) : \mathcal{X} \to \mathbb{R}^{\nu_1}$ based on state visits and a reward function $\boldsymbol{g} = (g_1, \ldots, g_{\nu_2}) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{\nu_2}$ based on transitions such that the run counts in a string of length n are obtained as additive functionals of the Markov chain, i.e. $\boldsymbol{S}_n = \sum_{m=0}^{n-1} \boldsymbol{f}(X_m)$ and $\boldsymbol{T}_n = \sum_{m=0}^{n-1} \boldsymbol{g}(X_m, X_{m+1})$. The components f_λ and g_λ are typically indicator functions and their choice depends on that of the Markov chain $\{X_n\}$ which is not unique.

2. When $\sum_{i \in \mathcal{X}} \pi_i f_\kappa(i) =: \mu_\kappa^f < \infty$, $\sum_{i \in \mathcal{X}} \pi_i f_\kappa^2(i) < \infty$ for $\kappa = 1, 2, \ldots, \nu_1$, and correspondingly $\sum_{(i,j) \in \mathcal{X} \times \mathcal{X}} \pi_i P_{ij} g_\lambda(i,j) := \mu_\lambda^g < \infty$, $\sum_{(i,j) \in \mathcal{X} \times \mathcal{X}} \pi_i P_{ij} g_\lambda^2(i,j) < \infty$, for $\lambda = 1, \ldots, \nu_2$, a Strong Law of Large Numbers and a multivariate CLT holds for $\{(\boldsymbol{S}_n, \boldsymbol{T}_n)\}_{n \in \mathbb{N}}$. Under these conditions, with $\boldsymbol{\mu}^f := (\mu_1^f, \ldots, \mu_{\nu_1}^f)$, $\boldsymbol{\mu}^g := (\mu_1^g, \ldots, \mu_{\nu_2}^g)$,

$$n^{-1/2} \left((\mathbf{S}_n, \boldsymbol{T}_n) - n(\boldsymbol{\mu}^f, \boldsymbol{\mu}^g) \right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{V}). \tag{2}$$

3. The means $\boldsymbol{\mu}^f$ and $\boldsymbol{\mu}^g$ are typically easy to compute in terms of the stationary distribution of $\{X_n\}$. On the other hand the $\nu \times \nu$ covariance matrix $\mathbf{V}$ in the multivariate CLT (2) is harder to determine and herein lies the main idea of the method which determines the covariance matrix by means of the potential matrix Z of the Markov chain $\{X_n\}$ and its stationary distribution π. The following expressions hold (see [2] and [24])

$$\lim_{n \to \infty} \frac{1}{n} \mathrm{Var}(S_n^\kappa) = 2 \sum_{i,j \in \mathcal{X}} f_\kappa(i) f_\kappa(j) \pi_i Z_{ij} - \left(\sum_{i \in \mathcal{X}} f_\kappa(i) \pi_i \right)^2 - \sum_{i,j \in \mathcal{X}} f_\kappa^2(i) \pi_i, \tag{3}$$
$$\kappa = 1, \ldots, \nu_1.$$

$$\lim_{n \to \infty} \frac{1}{n} \mathrm{Var}(T_n^\lambda) = 2 \sum_{l_1, l_2, k_1, k_2 \in \mathcal{X}} g_\lambda(l_1, l_2) g_\lambda(k_1, k_2) \pi_{l_1} P_{l_1 l_2} Z_{l_2 k_1} P_{k_1 k_2}$$
$$-3 \left(\sum_{l_1, l_2 \in \mathcal{X}} g_\lambda(l_1, l_2) \pi_{l_1} P_{l_1 l_2} \right)^2 + \sum_{l_1, l_2 \in \mathcal{X}} g_\lambda^2(l_1, l_2) \pi_{l_1} P_{l_1 l_2}, \qquad \lambda = 1, \ldots, \nu_2. \tag{4}$$

Corresponding expressions are given below for the covariances:

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(S_n^\kappa,S_n^\lambda) = \sum_{i,j\in\mathcal{X}} f_\kappa(i)f_\lambda(j)\left(\pi_i Z_{ij}+\pi_j Z_{ji}\right) - \sum_{i\in\mathcal{X}} f_\kappa(i)f_\lambda(i)\pi_i - \sum_{i\in\mathcal{X}} f_\kappa(i)\pi_i \sum_{j\in\mathcal{X}} f_\lambda(j)\pi_j, \qquad \kappa,\lambda=1,\ldots,\nu_1,\ \ \kappa\neq\lambda, \tag{5}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(T_n^\kappa,T_n^\lambda) = \sum_{l_1,l_2,k_1,k_2\in\mathcal{X}} g_\kappa(l_1,l_2)g_\lambda(k_1,k_2)P_{l_1l_2}P_{k_1k_2}\left(\pi_{l_1}Z_{l_2k_1}+\pi_{k_1}Z_{k_2l_1}\right)$$
$$+\sum_{l_1,l_2\in\mathcal{X}} g_\kappa(l_1,l_2)g_\lambda(l_1,l_2)\pi_{l_1}P_{l_1l_2} - 3\sum_{l_1,l_2\in\mathcal{X}} g_\kappa(l_1,l_2)\pi_{l_1}P_{l_1l_2}\sum_{k_1,k_2\in\mathcal{X}} g_\lambda(k_1,k_2)\pi_{k_1}P_{k_1k_2} \tag{6}$$
$$\kappa,\lambda=1,\ldots,\nu_2,\ \ \kappa\neq\lambda,$$

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(S_n^\kappa,T_n^\lambda) = \sum_{i,j,k\in\mathcal{X}} f_\kappa(i)g_\lambda(j,k)\left(\pi_i Z_{ij}+\pi_j Z_{ki}\right)P_{jk} \tag{7}$$
$$-2\sum_{i\in\mathcal{X}}\pi_i f_\kappa(i)\sum_{i,j\in\mathcal{X}} g_\lambda(i,j)\pi_i P_{ij}, \qquad \kappa=1,\ldots,\nu_1,\ \lambda=1,\ldots,\nu_2.$$

4. What remains is the determination of the stationary distribution π of the Markov chain $\{X_n\}$, which in many cases can obtained easily in closed form, and that of the potential matrix Z. In [24] this was obtained by means of its connection to mean transition times:

$$Z_{ij} = Z_{jj}-\pi_j\mathbb{E}_i\tau_j, \quad i\neq j, \tag{8}$$
$$Z_{ii} = \pi_i\mathbb{E}_\pi\tau_i. \tag{9}$$

where $\tau_i = \inf\{n>0; X_n = i\}$, $\mathbb{E}_i$ denotes conditional expectation given that $X_0 = i$, and $\mathbb{E}_\pi$ expectation assuming that X_0 is distributed according to the stationary distribution π (see [5]). The same approach is followed here since the mean transition times $E_i\tau_j$ can be obtained explicitly taking advantage of the special structure of the Markov chain considered. If this is not possible one can obtain the elements of the transition matrix numerically by means of the relationship $Z = (I-P+\ \ )^{-1}$ where is a square matrix having all rows equal to the stationary probability row vector π.

3 A Markov Chain for Multi-type Runs

Consider a sequence of trials $\{\xi_n; n\in\mathbb{N}\}$ with values in the finite set $\mathcal{S} = \{s_1,s_2,\ldots,s_d\}$ which forms an irreducible Markov chain with transition probability matrix P with elements p_{lr}, $l,r = 1,\ldots,d$. To avoid trivialities we will assume that $p_{ll}>0$ for all l. Denote the stationary probability distribution by η and the mean transition time from state r to state l by μ_{lr}. These can be obtained from the system $1+\sum_{r\neq t}p_{rt}\mu_{tl} = \mu_{rl}$, for $r,l = 1,\ldots,d$.

A special case of particular interest to which we shall occasionally refer is that of independent trials where the d symbols occur independently with probability $p_l>0$ for the occurrence of s_l, $l=1,\ldots,d$, with $\sum_{l=1}^d p_l$. In this case $p_{rl} = p_l$, the stationary distribution is given by $\eta_l = p_l$, and the mean transition times by $\mu_{rl} = p_l^{-1}$ for $l,r = 1,\ldots,d$.

3.1 A Markov chain for runs $G_{n,k}$, $M_{n,k}$, $J_{n,k}$, $S_{n,k}$

Construct a Markov chain $\{X_n; n \in \mathbb{N}\}$ with state space $\mathcal{U} = \{(r,i); r = 1,\ldots,d; i = 1,\ldots,k_r+1\}$. The labeling of the states reflects the fact that we visualize the state space as consisting of d branches, one for each type of source symbol. Thus state (r,i) signifies that the last symbol was s_r and that, looking back, i consecutive occurrences of s_r have occurred. As soon as a different symbol appears, say s_l, we move to state $(l,1)$ since this is the beginning of a run of s_l. By ordering the elements of the state space lexicographically, the transition probability matrix $\widetilde{P}$ of $\{X_n\}$ can be partitioned in blocks as follows

$$\widetilde{P} = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1d} \\ B_{21} & B_{22} & \cdots & B_{2d} \\ \vdots & & \ddots & \vdots \\ B_{d1} & B_{d2} & \cdots & B_{dd} \end{bmatrix}. \tag{10}$$

The diagonal blocks B_{rr} are square $(k_r+1)\times(k_r+1)$ matrices describing state transitions within the same run while the off-diagonal blocks B_{rl} are rectangular $(k_r+1)\times(k_l+1)$ matrices describing transitions from states (r,i), $i = 1,\ldots,k_r$, to state $(l,1)$ which occur whenever a run of s_r's is interrupted by the occurrence of an s_l. These blocks have the form

$$B_{rr} = \begin{bmatrix} 0 & p_{rr} & 0 & \cdots & 0 \\ 0 & 0 & p_{rr} & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & p_{rr} \\ 0 & 0 & 0 & \cdots & p_{rr} \end{bmatrix}, \qquad B_{rl} = \begin{bmatrix} p_{rl} & 0 & \cdots & 0 \\ p_{rl} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ p_{rl} & 0 & \cdots & 0 \end{bmatrix} \quad l \neq r. \tag{11}$$

All elements of B_{rr} are zero except for $[B_{rr}]_{i,i+1} = p_{rr}$ for $i = 1,2,\ldots,k_r$ and $[B_{rr}]_{k_r+1,k_r+1} = p_{rr}$. All elements of B_{rl} are zero except for $[B_{rr}]_{i,1} = p_{rr}$ for $i = 1,2,\ldots,k_r+1$. The Markov chain $\{X_n\}$ with the above transition probability matrix is irreducible, aperiodic, and positive recurrent in view of the assumptions on $\{\xi_n\}$. Its stationary distribution is given by

$$\begin{aligned} \pi_{l,i} &= \eta_l(1-p_{ll})p_{ll}^{i-1}, \qquad i = 1,2,\ldots,k_l, \\ \pi_{l,k_l+1} &= \eta_l p_{ll}^{k_l}. \end{aligned} \tag{12}$$

The total number of runs in n trials for each of the four different kinds of runs and for the d different type of symbols can be described in terms of additive functionals of the Markov chain $\{X_n; n \in \mathbb{N}\}$ as follows for $l = 1,\ldots,d$:

$$G_{n,k_l}^l = \sum_{m=0}^{n-1} \mathbf{1}\left(X_m = (l,k_l)\right), \tag{13}$$

$$M_{n,k_l}^l = \sum_{m=0}^{n-1} \mathbf{1}\left(X_m \in \{(l,k_l),(l,k_l+1)\}\right), \tag{14}$$

$$J_{n,k_l}^l = \sum_{m=0}^{n-1} \sum_{r\neq l} \mathbf{1}\left(X_m = (l,k_l), X_{m+1} = (r,1)\right), \tag{15}$$

$$S_{n,k_l}^l = \sum_{m=0}^{n-1} (k_l \mathbf{1}(X_m = (l,k_l) + \mathbf{1}(X_m = (l,k_l+1))\,. \tag{16}$$

3.2 Mean Transition Times and the Potential Matrix

Let $\tau_{l,j} := \min\{n > 0 : X_n = (l,j)\}$, $l \in \{1,\dots,d\}$, $j = 1,2,\dots,k_l+1$. We will denote the mean transition time between states of the runs chain by $m_{l,j}^{r,i} := \mathbb{E}_{r,i}\tau_{l,j}$. These can be easily expressed in terms of the mean transition times of the source chain as follows:

$$m_{l,j}^{r,i} = \mu_{rl} + \frac{p_{ll}^{-j+1}-1}{\eta_l(1-p_{ll})} \quad \text{for } l \neq r \text{ and} \quad m_{l,j}^{l,i} = \begin{cases} \dfrac{p_{ll}^{-j+1}-p_{ll}^{-i+1}}{\eta_l(1-p_{ll})}, & 1 \le i < j \le k_l+1, \\ \dfrac{p_{ll}^{-j+1}}{\eta_l(1-p_{ll})}, & 1 \le j < i \le k_l+1. \end{cases} \tag{17}$$

Using (9) and the above expressions for the mean transition times the diagonal potential matrix elements are given by

$$Z_{(l,i),(l,i)} = \begin{cases} 1-p_{ll}^{i-1}(1-\eta_l)\left(1+\eta_l\left(i-1-\sum_{r\neq l}\mu_{rl}\eta_r\right)\right), & \text{if } i = 1,\dots,k_d, \\ \frac{1}{1-p_{ll}}-p_{ll}^{k_l}\left(\frac{1}{1-p_{ll}}+\eta_l\left(k_l-1-\sum_{t\neq l}\eta_t\mu_{rl}\right)\right), & \text{if } i = k_d+1. \end{cases}$$

The off-diagonal elements are, for $j = 1,\dots,k_l$,

$$Z_{(r,i),(l,j)} = \begin{cases} \eta_l p_{ll}^{j-1}\left(1-(1-p_{ll})\left(j-1+\mu_{rl}-\sum_{t\neq l}\eta_t\mu_{tl}\right)\right), & \text{if } r \neq l, \\ -p_{ll}^{j-1}\left(1-\eta_l-p_{ll}^{-i+1}+\eta_l(1-p_{ll})\left(j-1-\sum_{t\neq l}\eta_t\mu_{tl}\right)\right), & \text{if } r = l,\ i < j, \\ -p_{ll}^{j-1}\left(1-\eta_l+\eta_l(1-p_{ll})\left(j-1-\sum_{t\neq l}\eta_t\mu_{tl}\right)\right), & \text{if } r = l,\ i > j, \end{cases}$$

and

$$Z_{(r,i),(l,k_l+1)} = \begin{cases} p_{ll}^{k_l}\left(\frac{p_{ll}^{-i+1}-1}{1-p_{ll}}-\eta_l\left(k_l-1-\sum_{t\neq l}\eta_t\mu_{tl}\right)\right), & \text{if } r = l, \\ -\eta_l p_{ll}^{k_l}\left(k_l-1+\mu_{rl}-\sum_{t\neq l}\eta_t\mu_{tl}\right), & \text{if } r \neq l. \end{cases}$$

3.3 A Runs Chain for $N_{n,\boldsymbol{k}}$

To deal with non-overlapping runs we construct the Markov chain $\{Y_n; n \in \mathbb{N}\}$ with state space $\mathcal{V} = \{(l,i); l = 1,\dots,d, i = 1,\dots,k_l\}$. Notice that each branch has one fewer state than in the corresponding state space of §3.1. The transition probability matrix has the same block structure as in (10). However, here B_{rr} is the $k_r \times k_r$ matrix

$$B_{rr} = \begin{bmatrix} & p_{rr} & & & \\ & & p_{rr} & & \\ & & & \ddots & \\ & & & & p_{rr} \\ p_{rr} & & & & \end{bmatrix}. \tag{18}$$

All elements of B_{rr} are zero except for $[B_{rr}]_{i,i+1} = p_{rr}$ for $i = 1, 2, \ldots, k_r - 1$ and $[B_{rr}]_{k_r,1} = p_{rr}$. The off-diagonal blocks B_{rl} are rectangular $k_r \times k_l$ matrices with all elements equal to zero except for the first column in which every element is equal to p_{rl}. The Markov chain $\{Y_n\}$ is irreducible, aperiodic, and positive recurrent. Its stationary distribution is given by

$$\pi_{l,i} = \frac{\eta_l(1-p_{ll})p_{ll}^{i-1}}{1-p_{ll}^{k_l}}, \qquad i = 1, 2, \ldots, k_l, \quad l = 1, \ldots, d. \tag{19}$$

The total number of non-overlapping s_l-runs of size k_l in n trials can be obtained from $\{Y_n\}$ as

$$N_{n,k_l}^{l} = \sum_{m=0}^{n-1} \mathbf{1}(Y_m = (l, k_l)). \tag{20}$$

3.4 Mean Hitting Times and Potential Matrix Elements

The mean transition times for $\{Y_n\}$ are given by

$$m_{l,j}^{r,i} = \mu_{rl} + \frac{p_{ll}^{-j+1} - 1}{\eta_l(1-p_{ll})} \quad \text{for } l \neq r \text{ and} \quad m_{l,j}^{l,i} = \begin{cases} \dfrac{p_{ll}^{-j+1} - p_{ll}^{-i+1}}{\eta_l(1-p_{ll})}, & 1 \le i < j \le k_l, \\ \dfrac{p_{ll}^{-j+1} - p_{ll}^{k_l-i+1}}{\eta_l(1-p_{ll})}, & 1 \le j < i \le k_l. \end{cases} \tag{21}$$

Thus the diagonal elements of the potential matrix are, for $l = 1, \ldots, d$, $i = 1, \cdots, k_l$,

$$Z_{(l,i),(l,i)} = \frac{1-(1-\eta_l)p_{ll}^{i-1}}{\eta_l(1-p_{ll})} + \frac{\eta_l(1-p_{ll})p_{ll}^{i-1}}{1-p_{ll}^{k_l}} \left(\frac{1-i(k_l+i-1)p_{ll}^{k_l}}{1-p_{ll}^{k_l}} + \sum_{t \neq l} \eta_t \mu_{tl} \right)$$

while the off-diagonal elements are, for $l, r = 1, \ldots, d$, $i = 1, \ldots, k_r$, $j = 1, \ldots, k_l$,

$$Z_{(r,i),(l,j)} = \begin{cases} \dfrac{1-(1-\eta_l)p_{ll}^{j-1}}{\eta_l(1-p_{ll})} + \dfrac{\eta_l(1-p_{ll})p_{ll}^{j-1}}{(1-p_{ll}^{k_l})^2} \left(-j + 1 + (k_l + j - 1)p_{ll}^{k_l} \right) \\ \qquad - \dfrac{1-p_{ll}^{j-1}}{1-p_{ll}^{k_l}} + \dfrac{\eta_l(1-p_{ll})p_{ll}^{j-1}}{1-p_{ll}^{k_l}} \left(\sum_{r \neq l} \mu_{rl}\eta_r - \mu_{rl} \right), & \text{if } r \neq l, \\ \dfrac{1-(1-\eta_l)p_{ll}^{j-1}}{\eta_l(1-p_{ll})} + \dfrac{(1-j)\eta_l(1-p_{ll})p_{ll}^{j-1} - \eta_l(1-k_l-j)(1-p_{ll})p_{ll}^{k_l+j-1}}{(1-p_{ll}^{k_l})^2} - \dfrac{1-p_{ll}^{j-i}}{1-p_{ll}^{k_l}} \\ \qquad + \dfrac{\eta_l(1-p_{ll})p_{ll}^{j-1}}{1-p_{ll}^{k_l}} \sum_{r \neq l} \mu_{rl}\eta_r, & \text{if } r = l,\ i < j, \\ \dfrac{1-(1-\eta_l)p_{ll}^{j-1}}{\eta_l(1-p_{ll})} + \dfrac{(1-j)\eta_l(1-p_{ll})p_{ll}^{j-1} - \eta_l(1-k_l-j)(1-p_{ll})p_{ll}^{k_l+j-1}}{(1-p_{ll}^{k_l})^2} \\ \qquad - \dfrac{1-p_{ll}^{k_l+j-i+1}}{1-p_{ll}^{k_l}} + \dfrac{\eta_l(1-p_{ll})p_{ll}^{j-1}}{1-p_{ll}^{k_l}} \sum_{r \neq l} \mu_{rl}\eta_r, & \text{if } r = l,\ i > j. \end{cases}$$

4 Main Results

Let $\{X_n\}$ be the Markov chain of §3.1. Using the representations (13)-(16) for the number of runs of various kinds in a string of length n and the CLT for additive functionals of Markov chains discussed in

§2 we obtain the multivariate CLT

$$\frac{1}{n^{1/2}}\left((\boldsymbol{G}_{n,\boldsymbol{k}},\boldsymbol{M}_{n,\boldsymbol{k}},\boldsymbol{J}_{n,\boldsymbol{k}},\boldsymbol{S}_{n,\boldsymbol{k}})-n(\boldsymbol{\mu}^G,\boldsymbol{\mu}^M,\boldsymbol{\mu}^J,\boldsymbol{\mu}^S\right)\xrightarrow{d}\mathcal{N}(0,\boldsymbol{V}). \tag{22}$$

The $4d$ components of the mean vector are given, for $l = 1,\ldots,d$ by

$$\begin{aligned}
\mu_l^G &= \mathbb{E}_\pi[\mathbf{1}(X_0=(l,k_l))] = \pi_{l,k_l} \\
\mu_l^M &= \mathbb{E}_\pi[\mathbf{1}(X_0\in\{(l,k_l),(l,k_l+1)\})] = \pi_{l,k_l}+\pi_{l,k_l+1} \\
\mu_l^J &= \mathbb{E}_\pi[\mathbf{1}(X_0=(l,k_l),X_1\in\{(r,1),r=1,\ldots,d,r\neq l\})] = (1-p_{ll})\pi_{l,k_l} \\
\mu_l^S &= \mathbb{E}_\pi[k_l\mathbf{1}(X_0=(l,k_l))+\mathbf{1}(X_0=(l,k_l+1))] = k_l\pi_{l,k_l}+\pi_{l,k_l+1},
\end{aligned} \tag{23}$$

where the stationary probabilities π_{l,k_l}, π_{l,k_l+1} are given by (12), (13). The elements of the covariance matrix $\boldsymbol{V}$ in (22) are given below. They are separated into two groups, covariances between run counts of the same kind but referring to runs of different symbols, and covariances between runs of different kind (and possibly of different symbols). They are expressed in terms of the stationary probabilities (12), (13) and the potential matrix whose elements are given in §3.2. In some of the cases expressions are given for the special case of independent trials. The expressions for the stationary distribution and the potential matrix elements in the case of independent trials are given in the Appendix.

Asymptotic Covariance of $\boldsymbol{G_{n,k}}$. Using (3) and (5) for additive functionals given by (13) we obtain

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(G_{n,k_l}^l) = 2\pi_{l,k_l}Z_{(l,k_l),(l,k_l)}-\pi_{l,k_l}(\pi_{l,k_l}+1), \tag{24}$$

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k_r}^r,G_{n,k_l}^l) = -\pi_{l,k_l}\pi_{r,k_r}+\pi_{l,k_l}Z_{(l,k_l),(r,k_r)}+\pi_{r,k_r}Z_{(r,k_r),(l,k_l)}. \tag{25}$$

In the case of independent trials

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(G_{n,k_l}^l) &= (1-p_l)p_l^{k_l}\left(1-(1-p_l)p_l^{k_l}(1+2k_l)\right), \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k_r}^r,G_{n,k_l}^l) &= p_r^{k_r}p_l^{k_l}\left(1-p_rp_l-(1-p_l)(1-p_r)(k_r+k_l)\right).
\end{aligned}$$

Asymptotic Covariance of $\boldsymbol{M_{n,k}}$. Again, from (3), (5), and (14) we obtain

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(M_{n,k_l}) &= 2\pi_{l,k_l}(Z_{(l,k_l),(l,k_l)}+Z_{(l,k_l),(l,k_l+1)})+2\pi_{l,k_l+1}(Z_{(l,k_l),(l,k_l+1)}+Z_{(l,k_l+1),(l,k_l)}) \\
&\quad -\pi_{l,k_l}(\pi_{l,k_l}+1)-\pi_{l,k_l+1}(\pi_{l,k_l+1}+1)-2\pi_{l,k_l}\pi_{l,k_l+1}, \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(M_{n,k_r},M_{n,k_l}) &= 2\pi_{l,k_l}(Z_{(l,k_l),(r,k_r)}+Z_{(l,k_l),(r,k_r+1)}) \\
&\quad +2\pi_{l,k_l+1}(Z_{(l,k_l+1),(r,k_r)}+Z_{(l,k_l+1),(r,k_r+1)})-(\pi_{l,k_l}+\pi_{l,k_l+1})(\pi_{r,k_r}+\pi_{r,k_r+1}) \\
&\quad +2\pi_{r,k_r}(Z_{(r,k_r),(l,k_l)}+Z_{(r,k_r),(l,k_l+1)})+2\pi_{r,k_r+1}(Z_{(r,k_r+1),(l,k_l)}+Z_{(r,k_r+1),(l,k_l+1)}).
\end{aligned}$$

Asymptotic Covariance of $J_{n,\boldsymbol{k}}$. From (4), (6), and (15) we obtain

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(J_{n,k_l}) &= 2\sum_{r,t\neq l}\pi_{l,k_l}p_{lr}p_{lt}Z_{(r,1)(l,k_l)} - 3\left(\sum_{r\neq l}\pi_{l,k_l}p_{lr}\right)^2 + \sum_{r\neq l}\pi_{l,k_l}p_{lr} \\
&= 2\pi_{l,k_l}(1-p_{ll})\sum_{r\neq l}p_{lr}Z_{(r,1)(l,k_l)} - 3\left(\pi_{l,k_l}(1-p_{ll})\right)^2 + \pi_{l,k_l}(1-p_{ll}), \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(J_{n,k_l}J_{n,k_r}) &= \pi_{l,k_l}(1-p_{rr})\sum_{t\neq l}p_{lt}Z_{(t,1)(r,k_l)} + \pi_{r,k_r}(1-p_{ll})\sum_{t\neq r}p_{rt}Z_{(t,1),(l,k_l)} \\
&\quad -3\pi_{l,k_l}\pi_{r,k_r}(1-p_{ll})(1-p_{rr}).
\end{aligned}$$

Asymptotic Covariance of $\boldsymbol{S}_{n,\boldsymbol{k}}$. From (3), (5), and (16) we obtain

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Var}(S_{n,k_l}) &= 2k_l\pi_{l,k_l}(Z_{(l,k_l),(l,k_l)} + Z_{(l,k_l),(l,k_l+1)}) + 2k_l\pi_{l,k_l+1}(Z_{(l,k_l),(l,k_l+1)} + Z_{(l,k_l+1),(l,k_l)}) \\
&\quad -k_l\pi_{l,k_l}(k_l\pi_{l,k_l}+1) - \pi_{l,k_l+1}(\pi_{l,k_l+1}+1) - 2k_l\pi_{l,k_l}\pi_{l,k_l+1}, \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(S_{n,k_r},s_{n,k_l}) &= 2\pi_{l,k_l}(k_lk_rZ_{(l,k_l),(r,k_r)} + k_lZ_{(l,k_l),(r,k_r+1)}) \\
&\quad +2\pi_{l,k_l+1}(k_rZ_{(l,k_l+1),(r,k_r)} + Z_{(l,k_l+1),(r,k_r+1)}) + 2\pi_{r,k_r}(k_lk_rZ_{(r,k_r),(l,k_l)} + k_rZ_{(r,k_r),(l,k_l+1)}) \\
&\quad +2\pi_{r,k_r+1}(k_lZ_{(r,k_r+1),(l,k_l)} + Z_{(r,k_r+1),(l,k_l+1)}) - (k_l\pi_{l,k_l}+\pi_{l,k_l+1})(k_r\pi_{r,k_r}+\pi_{r,k_r+1}).
\end{aligned}$$

Asymptotic Covariance for Different Kinds of Runs

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k_r},M_{n,k_l}) &= \pi_{l,k_l}Z_{(l,k_l),(r,k_r)} + \pi_{l,k_l+1}Z_{(l,k_l+1),(r,k_r)} + \pi_{r,k_r}(Z_{(r,k_r),(l,k_l)} + Z_{(r,k_r),(l,k_l+1)}) \\
&\quad -\pi_{r,k_r}(\pi_{l,k_l}+\pi_{l,k_l+1}) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(G_{n,k_r},S_{n,k_l}) &= k_l\pi_{l,k_l}Z_{(l,k_l),(r,k_r)} + \pi_{l,k_l+1}Z_{(l,k_l+1),(r,k_r)} + \pi_{r,k_r}(k_lZ_{(r,k_r),(l,k_l)} + Z_{(r,k_r),(l,k_l+1)}) \\
&\quad -\pi_{r,k_r}(k_l\pi_{l,k_l}+\pi_{l,k_l+1}) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(M_{n,k_r},S_{n,k_l}) &= 2k_l\pi_{l,k_l}(Z_{(l,k_l),(r,k_r)} + Z_{(l,k_l),(r,k_r+1)}) + 2\pi_{l,k_l+1}(Z_{(l,k_l+1),(r,k_r)} + Z_{(l,k_l+1),(r,k_r+1)}) \\
&\quad +2\pi_{r,k_r}(k_lZ_{(r,k_r),(l,k_l)} + Z_{(r,k_r),(l,k_l+1)}) + 2\pi_{r,k_r+1}(k_lZ_{(r,k_r+1),(l,k_l)} + Z_{(r,k_r+1),(l,k_l+1)}) \\
&\quad -(k_l\pi_{l,k_l}+\pi_{l,k_l+1})(\pi_{r,k_r}+\pi_{r,k_r+1}) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(J_{n,k_r},G_{n,k_l}) &= \sum_{t\neq r}p_{rt}\left(\pi_{r,k_r}Z_{(t,1)(l,k_l)} + \pi_{l,k_l}Z_{(l,k_l)(r,k_r)} - 2\pi_{l,k_l}\pi_{r,k_r}\right)
\end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(J_{n,k_r},M_{n,k_l}) &= \sum_{t\neq r}p_{rt}\left(\pi_{r,k_r}Z_{(t,1)(l,k_l)} + \pi_{l,k_l}Z_{(l,k_l)(r,k_r)}\right) \\
&\quad + \sum_{t\neq r}p_{rt}\left(\pi_{r,k_r}Z_{(t,1)(l,k_l+1)} + \pi_{l,k_l+1}Z_{(l,k_l+1)(r,k_r)}\right) - 2\sum_{t\neq r}p_{rt}\pi_{r,k_r}(\pi_{l,k_l}+\pi_{l,k_l+1}) \\
\lim_{n\to\infty}\frac{1}{n}\mathsf{Cov}(J_{n,k_r},S_{n,k_l}) &= \sum_{t\neq r}p_{rt}\left(\pi_{r,k_r}k_lZ_{(t,1)(l,k_l)} + \pi_{l,k_l}k_lZ_{(l,k_l)(r,k_r)}\right) \\
&\quad + \sum_{t\neq r}p_{rt}\left(\pi_{r,k_r}Z_{(t,1)(l,k_l+1)} + \pi_{l,k_l+1}Z_{(l,k_l+1)(r,k_r)}\right) - 2\sum_{t\neq r}p_{rt}\pi_{r,k_r}(k_l\pi_{l,k_l}+\pi_{l,k_l+1})
\end{aligned}$$

4.1 A CLT for $\boldsymbol{N}_{n,\boldsymbol{k}}$.

Here we deal with the number of exact runs of the symbols s_l, $l = 1, \ldots, d$ using the Markov chain $\{Y_n\}$ of §3.3. It holds that $n^{-1/2}\left(\boldsymbol{N}_{n,\boldsymbol{k}} - n\boldsymbol{\mu}^N\right) \xrightarrow{d} \mathcal{N}(0, V^N)$ where $\boldsymbol{\mu}^N = \mathbb{E}_\pi(\mathbf{1}(Y_0 = (l, k_l)) = \pi_{l,k_l}$ (given by (19)). The covariance matrix V^N has elements V_{rl} equal to the asymptotic variances and covariances

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(N_{n,k_l}) &= 2\pi_{l,k_l} Z_{(l,k_l),(l,k_l)} - \pi_{l,k_l}(\pi_{l,k_l} + 1), \\
\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(N_{n,k_r}, N_{n,k_l}) &= -\pi_{l,k_l}\pi_{r,k_r} + \pi_{l,k_l} Z_{(l,k_l),(r,k_r)} + \pi_{r,k_r} Z_{(r,k_r),(l,k_l)} \ .
\end{aligned}$$

In the above expressions the elements of the potential matrix Z are the ones obtained in §3.4.

5 Applications

Runs of any length Denote by T_n^l the number of runs of any length of the symbol s_l in a string of length n. Clearly, using the Markov chain of subsection 3.1 with $k_l = 1$, $l = 1, \ldots, d$, we have $T_n^l = G_{n,1}^l$ and for the vector statistic $\boldsymbol{T}_n := (T_n^1, \ldots, T_n^d)$ the results of the previous section apply. Thus $n^{-1/2}(T_n - n\boldsymbol{\mu}^T)$ converges in distribution to a multivariate normal. The mean vector is given by $\mu_l^T = \pi_{l,1}$. The covariance matrix is given by

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(T_{n,l}^l) &= 2\pi_{l,1} Z_{(l,1),(l,1)} - \pi_{l,1}(\pi_{l,1} + 1) \\
&= \eta_l(1 - p_{ll})[\eta_l(1 + p_{ll}) - 1] + 2\eta_l^2(1 - \eta_l)(1 - p_{ll}) \sum_{t\neq l} \eta_l \mu_{tl}, \\
\lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(T_{n,r}^r, T_{n,l}^l) &= -\pi_{l,1}\pi_{r,1} + \pi_{l,1} Z_{(l,1),(r,1)} + \pi_{r,1} Z_{(r,1),(l,1)} \\
&= \eta_l \eta_r (1 - p_{ll})(1 - p_{ll}) \left[2 + \mu_{lr} + \mu_{rl} - \sum_{t=1}^{d} \eta_t(\mu_{tl} + \mu_{tr})\right].
\end{aligned}$$

In the case of independent trials we have

$$\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(T_{n,l}^l) = p_l(1 - p_l)[1 - 3p_l(1 - p_l)], \quad \lim_{n\to\infty} \frac{1}{n}\mathsf{Cov}(T_{n,r}^r, T_{n,l}^l) = 2p_r p_l(1 - p_r)(1 - p_l).$$

Total number of runs. Summing up the random variables G_{n,k_l}^l over $l = 1, \ldots, d$ we obtain the statistic R_n denoting the number of runs of all symbols exceeding in length k_l for s_l. Then, using the Markov chain of subsection 3.1,

$$R_n = \sum_{m=0}^{n-1} \sum_{l=1}^{d} \mathbf{1}(X_m = (l, k_l)).$$

We have again a CLT with mean $\mu^R = \sum_{l=l}^{d} \pi_{l,k_l}$ and asymptotic variance

$$\lim_{n\to\infty} \frac{1}{n}\mathsf{Var}(R_n) = 2\sum_{l=1}^{d}\sum_{r=1}^{d} \pi_{l,k_l} Z_{(l,k_l)(r,k_r)} - \left(\sum_{l=1}^{d} \pi_{l,k_l}\right)^2 - \sum_{l=1}^{d} \pi_{l,k_l}. \tag{26}$$

Reliability of general multi-consecutive systems. The statistical analysis of runs has immediate applications to the so-called multi-consecutive systems in reliability theory. (For a review of the literature on such systems see [19].) The consecutive k-out-of-n : F system, as originally defined, consists of n components ordered on a line, each being independently either defective or non-defective. Such a system fails if and only if there are at least k consecutive failed components. A generalization, the m-consecutive-k-out-of-n : F system was introduced by Griffith [14]. In this case the system fails when m non-overlapping runs of k consecutive failed components occur. The exact reliability of this system has been studied by a number of authors. In particular Godbole [13] derived Poisson approximations for its reliability under the assumption of Markovian dependence using the Stein-Chen method. A normal approximation for the reliability of this system is given in Makri and Psillakis [22] in terms of $\mathbb{P}(G_{n,k} < m)$. Agarwal et al. [1] studied a variation of the m-consecutive-k-out-of-n : F system where the components exhibit markovian dependence.

Boutsikas and Koutras [4] considered components with d *different failure modes* and studied the consecutive k_1,k_2,...,k_d-out-of-n:MFM system (Multi-Failure Mode). It consists of n linearly arranged components and enters failure mode l whenever at least k_l consecutive components are failed in mode l, $l = 1, 2, \ldots, d$. Reliability bounds based on compound Poisson approximations for the consecutive k_1, k_2,...,k_d-out-of-n:MFM system with Markov-dependent components were derived by Chryssaphinou and Vaggelatou [6] using the Stein-Chen method.

We extent the concept of consecutive $k_1, k_2, \ldots, k_d$-out-of-n:MFM systems by assuming that the system consists of n linearly arranged components and enters failure mode l whenever m_l runs of k_l consecutive components fail in mode $l = 1, 2, \ldots, d$. We term this system a consecutive (k_1, m_1), (k_2, m_2), ..., (k_d, m_d)-out-of-n:MFM system and assume that the state of the components has Markovian dependence. Depending on the system it may be natural to count the number of runs for mode-l failures either as overlapping or non-overlapping. If s_0 stands for normal operation and s_l, $l = 1, \ldots, d$ for the d different failure modes then the results of the previous section can be used to obtain normal approximations for the reliability of such systems. For instance, for non-overlapping runs the reliability would be $\mathbb{P}(N_{n,k_1} < m_1, N_{n,k_2} < m_2, \ldots, N_{n,k_d} < m_d)$, while for overlapping runs it would be $\mathbb{P}(M_{n,k_1} < m_1, M_{n,k_2} < m_2, \ldots, M_{n,k_d} < m_d)$. The results of the previous section provide directly normal approximations for the reliability of such systems.

Molecular biology. It has been pointed out in Lou [21] and Fu et al. [12] that $S_{n,k}$ is useful in problems that arise in molecular biology and in particular in studying tandem repeats among DNA sequences. In this problem a binary sequence of successes and failures is derived by aligning two adjacent DNA sequences one on top of the other. $S_{n,k}$ denotes the sum of the matches in matching runs of length k or larger in a sequence of length n. The normal approximation for $S_{n,k}$ is given in Makri and Psillakis [22] for Bernoulli trials and by Fu et al. [12] for Markov dependent trials. When the matching process has markovian correlation and can be described by a two state chain with transition probability matrix

$$\begin{bmatrix} q_0 & p_0 \\ q & p \end{bmatrix}$$

a CLT holds for $S_{n,k}$. The asymptotic mean and variance are given by $\mu = \frac{p_0 p^{k-1}}{q+p_0}$ and

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(S_{n,k}) = \frac{k^2 q p_0 p^{k-1} + p_0 p^k (2k-1)}{q+p_0} + 2\frac{p_0 p^k}{q(q+p_0)} - 2\frac{p^{2k} p_0^2 (kq+p)}{q(q+p_0)^2}$$
$$- \frac{(1+2k)p^{2k} p_0^2 + (4q-p)k^2 q p_0^2 p^{2k-2} + 2k^3 q^2 p_0^2 p^{2k-2}}{(q+p_0)^2} - 2\frac{p_0^2 p^{2k} + 4kqp_0^2 p^{2k-1} + 2k^2 q^2 p_0^2 p^{2k-2}}{(q+p_0)^3}.$$

6 Hidden Markov Sources

A more general model for the multi-type source involves an underlying Markov chain $\{\zeta_n\}$, $n = 0, 1, 2, \ldots$ with state space $\mathcal{S}$ partitioned into subsets $\mathcal{S}_l$, $l = 1, \ldots, d$, so that $\mathcal{S}_l \cap \mathcal{S}_r = \emptyset$ for $l \neq r$ and $\cup_{l=1}^{d} \mathcal{S}_l = \mathcal{S}$. Let $m_l := |\mathcal{S}_l|$ denote the cardinality of $\mathcal{S}$ and label the elements of $\mathcal{S}$ as (q, l) where $q = 1, 2, \ldots, m_l$ and $l = 1, 2, \ldots, d$. Ordering the states of $\mathcal{S}$ lexicographically we obtain a partitioned matrix

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1d} \\ P_{21} & P_{22} & \cdots & P_{2d} \\ \vdots & \vdots & & \vdots \\ P_{d1} & P_{d2} & \cdots & P_{dd} \end{bmatrix} \tag{27}$$

where P_{rr} is the $m_r \times m_r$ submatrix containing the transition probabilities between states within class $\mathcal{S}_r$, while P_{rl}, $(l \neq r)$ is the $m_r \times m_l$ rectangular submatrix P containing the transition probabilities from states in $\mathcal{S}_r$ to states in $\mathcal{S}_l$. For $n \in \mathbb{N}$ let $\xi_n = l$ when $\zeta_n \in \mathcal{S}_l$, $l = 1, \ldots, d$. We will assume that the matrix P above is irreducible and, for simplicity, aperiodic. We will further assume that none of the diagonal matrices P_{rr}, $r = 1, \ldots, d$ is a zero matrix. We will denote the stationary distribution of $\{\zeta_n\}$ by the row vector $[\eta_1, \ldots, \eta_l, \ldots, \eta_d]$ in block form, where η_l is an m_l-dimensional row vector.

The process $\{\xi_n; n \in \mathbb{N}\}$ defined above is a sequence of *dependent multi-type trials* and we will use the approach of the previous sections in order to obtain a multivariate CLT for the number of success runs of various types.

Suppose we are interested in l-runs of size k_l for $l = 1, \ldots, d$. Construct a Markov chain $\{X_n\}$, $n \in \mathbb{N}$, on a larger state space, $\mathcal{X} := \{(l, i, q) : l = 1, \ldots, d, i = 1, \ldots, k_l + 1, q = 1, \ldots, m_l\}$. Order the states of $\mathcal{X}$ lexicographically. The transition probability matrix of $\{X_n\}$ with this ordering can be written in block form as in (10). Block B_{rr} is an $m_r(k_r + 1) \times m_r(k_r + 1)$ square matrix while B_{rl} $(r \neq l)$ an $m_r(k_r + 1) \times m_l(k_l + 1)$ rectangular matrix given respectively by

$$B_{rr} = \begin{bmatrix} O & P_{rr} & O & \cdots & O \\ O & O & P_{rr} & \cdots & O \\ & & & \ddots & \\ O & O & O & \cdots & P_{rr} \\ O & O & O & \cdots & P_{rr} \end{bmatrix}, \quad B_{rl} = \begin{bmatrix} P_{rl} & O & \cdots & O \\ P_{rl} & O & \cdots & O \\ \vdots & \vdots & & \vdots \\ P_{rl} & O & \cdots & O \end{bmatrix}$$

where O is a matrix of zero elements of appropriate dimensions (an $m_r \times m_r$ square matrix with zero elements for B_{rr} and an $m_r \times m_l$ rectangular submatrix of zeros for B_{rl}).

Under the assumptions on the matrix P in (27) $\{X_n\}$ is irreducible and aperiodic and we will denote its stationary distribution by $\pi(l, i, q)$, $l = 1, \ldots, L$, $i = 1, \ldots, k_l + 1$, $q = 1, \ldots, m_l$. It is given by

$$\pi(l, i) = \eta_l (I - P_{ll}) P_{ll}^{i-1}, \quad i = 1, \ldots, k_l, \tag{28}$$

$$\pi(l, k_l + 1) = \eta_l P_{ll}^{k_l}, \tag{29}$$

where $\pi(l, i)$ is a row vector of dimension m_l.

The number of the various kinds of runs in a string of length n is then given by

$$\begin{aligned}
M^l_{n,k_l} &= \sum_{j=0}^{n-1}\sum_{q=1}^{m_l} \mathbf{1}(X_j = (l,k_l,q)) + \mathbf{1}(X_j = (l,k_l+1,q)),\\
S^l_{n,k_l} &= \sum_{j=0}^{n-1}\sum_{q=1}^{m_l} k_l\mathbf{1}(X_j = (l,k_i,q)) + \mathbf{1}(X_j = (l,k_l+1,q)),\\
G^l_{n,k_l} &= \sum_{j=0}^{n-1}\sum_{q=1}^{m_l} \mathbf{1}(X_j = (l,k_l,q)),\\
J^l_{n,k_l} &= \sum_{j=0}^{n-1}\sum_{q=1}^{m_l}\sum_{\substack{r=1\\ r\neq l}}^{d}\sum_{t=1}^{m_r} \mathbf{1}(X_j = (l,k_l,q), X_{j+1} = (r,1,t)).
\end{aligned}$$

7 Appendix: Independent Multi-trials

A special case of particular interest is when the source produces symbols s_l independently with probabilities p_l, $l = 1,\ldots,d$. Then the stationary distribution of the Markov chain of subsection 3.1 is given, for $l = 1,\ldots,d$, by

$$\begin{aligned}
\pi_{l,i} &= (1-p_l)p_l^i, \quad i = 1,\ldots,k_l,\\
\pi_{l,k_l+1} &= p_l^{k_l+1}.
\end{aligned}$$

The diagonal potential matrix elements are given by

$$Z_{(l,i),(l,i)} = \begin{cases} 1 - i(1-p_l)p_l^i, & \text{if } i = 1,\ldots,k_l\\ \frac{1-p_l^{k_l+1}}{1-p_l} - p_l^{k_l+1}k_l, & \text{if } i = k_l+1. \end{cases}$$

The off-diagonal elements are, for $j = 1,\ldots,k_l$,

$$Z_{(r,i),(l,j)} = \begin{cases} p_l^j(1-(1-p_l)j), & \text{if } r \neq l,\\ p_l^{j-i} - j(1-p_l)p_l^j, & \text{if } r = l,\ i < j,\\ -j(1-p_l)p_l^j, & \text{if } r = l,\ i > j, \end{cases}$$

and

$$Z_{(r,i),(l,k_l+1)} = \begin{cases} p_l^{k_l+1}\frac{p_l^{-i}-1}{1-p_l} - k_l p_l^{k_l+1}, & \text{if } r = l,\\ -k_l p_l^{k_l+1}, & \text{if } r \neq l. \end{cases}$$

References

[1] Agarwal, M., Sen K. and Mohan P. (2007). GERT analysis of m-consecutive-k-out-of-n systems. *IEEE Trans. Reliab.* **56**, 2634.

[2] Asmussen, S. (2003). *Applied Probability and Queues.* 2nd ed. Springer Verlag.

[3] Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, Wiley, New York.

[4] Boutsikas, M.V. and Koutras, M.V. (2002). On a class of multiple failure mode systems. *Naval Research Logistics* **49**, 167-185.

[5] Brémaud, P. (1997). *Markov Chains*, Springer Verlag.

[6] Chryssaphinou, O. and Vagelatou, E. (2002). Compound Poisson approximation for multiple runs in a Markov chain. *Annals of the Institute of Statistical Mathematics* **54**, 411-424.

[7] Doi, M. and Yamamoto, E.(1998). On the joint distribution of runs in a sequence of multi-state trials. *Statistics and Probability Letters* **39**, 133-141.

[8] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications vol. 1*, 3rd edition. John Wiley, New York.

[9] Fu, J.C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica* **6**, 957-974.

[10] Fu, J. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and its Applications,* World Scientific, Singapore.

[11] Fu, J. and 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.

[12] Fu, J., Lou, W.Y.W., Bai, Z.D. and 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.

[13] Godbole, A.P. (1993). Approximate reliabilities of m-consecutive-k-out-of-n: failure systems. *Stat. Sinica* **3**, 321-327.

[14] Griffith, W.S. (1986). On consecutive-k-out-of-n failure systems and their generalizations. *Basu A.P. (ed) Reliability and quality control* Elsevier, North Holland, 157-165.

[15] Gibbons, J.D. (1971). *Nonparametric Statistical Inference,* McGraw-Hill, New York.

[16] Han, Q. and Aki, S. (1999). Joint distributions of runs in a sequence of multi-state trials. *Annals of the Institute of Statistical Mathematics* **51** 419-447.

[17] Inoue K. and Aki S. (2007) Joint distributions of numbers of runs of specified lengths in a sequence of Markov dependent multistate trials. *Annals of the Institute of Statistical Mathematics* **54**, 719-730.

[18] Koutras, M. V., Alexandrou, V. A. (1997). Non-parametric randomness tests based on success runs of fixed length. *Statistics and Probability Letters* **32**, 393-404.

[19] Kuo, W. and Zuo, M. (2003). *Optimal reliability modeling: princi- ples and applications*. Wiley, New Jersey.

[20] Ling, K. D. (1988). On binomial distributions of order k. *Statistics and Probability Letters* **6**, 247-250.

[21] Lou, W. (2003). The exact distribution of the k-tuple statistic for sequence homology. *Statistics and Probability Letters* **61**, 51-59.

[22] Makri, F.S. and Psillakis, Z.M. (2009). On runs of length exceeding a threshold: normal approximation. *Statistical Papers*, Springer.

[23] Mood, A.D. (1940). The distribution theory of runs. *Annals of Mathematical Statistics* **11**, 367-392.

[24] Mytalas, G.C. and Zazanis M.A. (2013). Central Limit Theorem Approximations for the Number of Runs in Markov-Dependent Binary Sequences. *Journal of Statistical Planning and Inference*, **143** 321-333.

[25] Shinde, R. L. Kotwal, K. S. (2006). On the joint distribution of runs in the sequence of Markov-dependent multi-state trials. *Statistics and Probability Letters* **76**, 1065-1074.