Article

One Dimensional Discrete Scan Statistics for Dependent Models and Some Related Problems

Alexandru Amarioarei ^{1,2} and Cristian Preda ^{3,4,5,6,*}¹ Faculty of Mathematics and Computer Science, University of Bucharest, 010014 Bucharest, Romania; alexandru.amarioarei@fmi.unibuc.ro² National Institute of Research and Development for Biological Sciences, 060031 Bucharest, Romania³ Laboratoire de Mathématiques Paul Painlevé, University of Lille, 59655 Villeneuve d'Ascq, France⁴ Biostatistics Department, Delegation for Clinical Research and Innovation, Lille Catholic Hospitals, GHICL, 59462 Lomme, France⁵ Institute of Statistics and Applied Mathematics of the Romanian Academy, 050711 Bucharest, Romania⁶ Inria Lille Nord-Europe, MODAL, 59655 Villeneuve d'Ascq, France

* Correspondence: cristian.preda@univ-lille.fr

Received: 18 March 2020; Accepted: 9 April 2020; Published: 13 April 2020

Abstract: The one dimensional discrete scan statistic is considered over sequences of random variables generated by block factor dependence models. Viewed as a maximum of an 1-dependent stationary sequence, the scan statistics distribution is approximated with accuracy and sharp bounds are provided. The longest increasing run statistics is related to the scan statistics and its distribution is studied. The moving average process is a particular case of block factor and the distribution of the associated scan statistics is approximated. Numerical results are presented.

Keywords: scan statistics; 1-dependence; runs; approximation; simulation

1. Introduction

There are many situations when an investigator observes an accumulation of events of interest and wants to decide if such a realisation is due to hazard or not. These types of problems belong to the class of cluster detection problems, where the basic idea is to identify regions that are *unexpected* or *anomalous* with respect to the distribution of events. Depending on the application domain, these *anomalous* agglomeration of events can correspond to a diversity of phenomena—for example one may want to find clusters of stars, deposits of precious metals, outbreaks of disease, minefield detections, defectuous batches of pieces and many other possibilities. If such an observed accumulation of events exceeds a preassigned threshold, usually determined from a specified significance level corresponding to a *normal* situation (the null hypothesis), then it is legitimate to say that we have an unexpected cluster and proper measures has to be taken accordingly.

Searching for unusual clusters of events is of great importance in many scientific and technological fields, including DNA sequence analysis ([1,2]), brain imaging ([3]), target detection in sensors networks ([4,5]), astronomy ([6,7]), reliability theory and quality control ([8]) among many other domains. One of the tools used by practitioners to decide on the unusualness of such agglomeration of events is the *scan statistics*. Basically, the tests based on scan statistics are looking for events that are clustered amongst a background of those that are sporadic.

Let $2 \leq m \leq T$ be two positive integers and $X_1, \dots, X_T$ be a sequence of independent and identically distributed random variables with the common distribution F_0 . The *one dimensional discrete scan statistics* is defined as

$$S_m(T) = \max_{1 \leq i \leq T-m+1} W_i, \quad (1)$$

where the random variables W_i are the moving sums of length m given by

$$W_i = \sum_{j=i}^{i+m-1} X_j. \quad (2)$$

Usually, the statistical tests based on the one dimensional discrete scan statistics are employed when one wants to detect a local change in the signal within a sequence of T observations via testing the null hypothesis of uniformity, H_0 , against a cluster alternative, H_1 (see References [9,10]). Under H_0 , the random observations $X_1, \dots, X_T$ are i.i.d. distributed as F_0 , while under the alternative hypothesis, there exists a location $1 \leq i_0 \leq T - m + 1$ where $X_i, i \in \{i_0, \dots, i_0 + m - 1\}$, are distributed according to $F_1 \neq F_0$ and outside this region X_i are distributed as F_0 .

We observe that whenever $S_m(T)$ exceeds the threshold τ , where the value of τ is computed based on the relation $\mathbb{P}_{H_0}(S_m(T) \geq \tau) = \alpha$ and α is a preassigned significance level of the testing procedure, the generalized likelihood ratio test rejects the null hypothesis in the favor of the clustering alternative (see Reference [9]). It is interesting to note that most of the research has been done for F_0 being binomial, Poisson or normal distribution (see References [9–13]). More recently, Reference [14] proposed a testing procedure based on one-dimensional scan statistic for geometric and negative binomial distributions.

There are three main approaches used for investigating the exact distribution of the one dimensional discrete scan statistics—the combinatorial methods ([12,15]), the Markov chain imbedding technique ([16,17]) and the conditional probability generating function method ([18,19]). Due to the high complexity and the limited range of application of the exact formulas, a considerable number of approximations and bounds have been developed for the estimation of the distribution of the one dimensional discrete scan statistics, for example, References [9,12,13,20]. A full treatment of these results is presented in References [10,11].

Even if in general the X_i 's are supposed to be i.i.d distributed, there are applications, such as detecting similarities between DNA sequences, where the X_i 's are not independent ([21]). In order to evaluate the effect of dependence, the alternative model is in many cases a Markov chain the whole dependence structure of which depends only on the joint distribution of two consecutive random variables.

In this work we introduce dependence models based on block-factors obtained from i.i.d. sequences in the context of the one dimensional discrete scan statistics. We derive approximations and their corresponding errors with application to the longest increasing run distribution and the moving average process.

The paper is structured as follows. In Section 2 we introduce the block factor model and present the approximation technique for the distribution of the scan statistics under this model. As a particular block factor model, the distribution of the length of the longest increasing run in a trial of i.i.d real random variables and the moving average processes are related to the scan statistics distribution in Section 3. Numerical results based on simulations illustrate the accuracy of the approximation. Concluding remarks end the paper.

2. One Dimensional Scan Statistics for Block-Factor Dependence Model

Most of the research devoted to the one dimensional discrete scan statistic considers the independent and identically distributed model for the random variables that generate the sequence which is to be scanned. In this section, we define a dependence structure for the underlying random sequence based on a block-factor type model.

2.1. The Block-Factor Dependence Model

Let us recall (see also Reference [22]) that the sequence $\{X_i\}_{i \geq 1}$ of random variables with state space S_W is said to be k block-factor of the sequence $\{Y_i\}_{i \geq 1}$ with state space S_Y if there is a measurable function $f : S_Y^k \rightarrow S_W$ such that

$$X_i = f(Y_i, Y_{i+1}, \dots, Y_{i+k-1}), \forall i \geq 1. \quad (3)$$

The Figure 1 presents the sequence $\{X_i\}_{i=1, \dots, T}$ of length T obtained as a k block-factor from a sequence $\{Y_i\}_{i=1, \dots, \tilde{T}}$ of length $\tilde{T} = T + k - 1$ throughout some function f .

Figure 1. The block factor model.

As an example of block-factor model, in Reference [23], the authors consider an i.i.d. sequence $\{Y_n\}_{n \geq 1}$ of standard normal distributed random variables and the 2 block-factor defined by

$$X_i = aY_i + bY_{i+1}, i \geq 1, \text{ for } f(x, y) = ax + by, a, b \in \mathbb{R}.$$

Therefore, due to the overlapping structure of X_i and X_{i+1} , they obtain a Gaussian stationary process $\{X_i\}_{i \geq 1}$ with some correlation structure for which the scan statistics distribution is studied.

More generally, let observe that if a sequence $\{X_i\}_{i \geq 1}$ of random variables is a k block-factor, then the sequence is $(k - 1)$ -dependent. Recall that a sequence $\{X_i\}_{i \geq 1}$ is m -dependent with $m \geq 1$ (see Reference [22]), if for any $h \geq 1$ the σ -fields generated by $\{X_1, \dots, X_h\}$ and $\{X_{h+m+1}, \dots\}$ are independent.

2.2. Scan Statistics Viewed as Maximum of 1-Dependent Sequence

Let $\{X_1, \dots, X_T\}$ be a k -block factor of the i.i.d. sequence $\{Y_1, \dots, Y_{\tilde{T}}\}$, where $\tilde{T}_1 \geq k$ and $T = \tilde{T} - k + 1$, and $S_m(T)$ be the scan statistics associated to the sequence $\{X_i\}_{i=1, \dots, T}$ as defined in (1) for some scanning window of length m , $1 \leq m \leq T$.

Put $T = (L - 1)(m + k - 2) + m - 1$ for some integer $L \geq 1$ and define, for each $j \in \{1, \dots, L - 1\}$, the random variables

$$Z_j = \max_{(j-1)(m+k-2)+1 \leq i \leq j(m+k-2)} W_i, \quad (4)$$

where $W_i = \sum_{s=i}^{i+m-1} X_s$. That is, for each $j \in \{1, \dots, L - 1\}$, Z_j is the scan statistic associated to the sequence of length $2m + k - 3$, $\{X_{(j-1)(m+k-2)+1}, \dots, X_{j(m+k-2)+m-1}\}$. An illustration of the construction of variables Z_j is presented in Figure 2 for $L = 5$ and $k = 1$.

Figure 2. Construction of Z_j .

Then, $\{Z_j\}_{j=1,\dots,L-1}$ is 1-dependent and stationary and we have that

$$S(m, T) = \max_{1 \leq j \leq L-1} Z_j. \quad (5)$$

Thus, for any block factor model obtained from an i.i.d sequence, the distribution of the associated scan statistics is the distribution of the maximum of some 1-dependent stationary sequence.

2.3. Approximation

In References [24,25] the authors extended the approximation results obtained in Reference [26] for the distribution of the maximum of a 1-dependent stationary sequence. The main results is stated in the following theorem.

Let $\{Z_j\}_{j \geq 1}$ be a strictly stationary 1-dependent sequence of random variables and for $x < \sup\{u | \mathbb{P}(Z_1 \leq u) < 1\}$, let

$$1q_n = q_n(x) = \mathbb{P}(\max(Z_1, \dots, Z_n) \leq x). \quad (6)$$

Theorem 1. For all x such that $q_1(x) \geq 1 - \alpha \geq 0.9$, the following approximation formula holds:

$$\left| q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n} \right| \leq nF(\alpha, n)(1 - q_1)^2 \quad (7)$$

with

$$F(\alpha, n) = 1 + \frac{3}{n} + \left[\frac{\Gamma(\alpha)}{n} + K(\alpha) \right] (1 - q_1) \quad (8)$$

where $\Gamma(\alpha) = L(\alpha) + E(\alpha)$,

$$K(\alpha) = \frac{\frac{11-3\alpha}{(1-\alpha)^2} + 2l(1+3\alpha) \frac{2+3l\alpha-\alpha(2-l\alpha)(1+l\alpha)^2}{[1-\alpha(1+l\alpha)^2]^3}}{1 - \frac{2\alpha(1+l\alpha)}{[1-\alpha(1+l\alpha)^2]^2}} \quad (9)$$

$$L(\alpha) = 3K(\alpha)(1 + \alpha + 3\alpha^2)[1 + \alpha + 3\alpha^2 + K(\alpha)\alpha^3] + \alpha^6 K^3(\alpha) + 9\alpha(4 + 3\alpha + 3\alpha^2) + 55.1 \quad (10)$$

$$E(\alpha) = \frac{\eta^5 [1 + (1 - 2\alpha)\eta]^4 [1 + \alpha(\eta - 2)] [1 + \eta + (1 - 3\alpha)\eta^2]}{2(1 - \alpha\eta^2)^4 [(1 - \alpha\eta^2)^2 - \alpha\eta^2(1 + \eta - 2\alpha\eta)^2]} \quad (11)$$

and where $\eta = 1 + l\alpha$ with $l = l(\alpha) = t_2^3(\alpha) + \varepsilon$, for arbitrarily small $\varepsilon > 0$, and $t_2(\alpha)$ the second root in magnitude of the equation $\alpha t^3 - t + 1 = 0$.

The evaluation of the functions K and Γ for some selected values of α is presented in Table 1. These values allow to compute directly the error bound approximation in (7).

Table 1. Selected values for l , K and Γ functions in Theorem 1 for $\varepsilon = 10^{-6}$.

α	$l(\alpha)$	$K(\alpha)$	$\Gamma(\alpha)$
0.1	1.5347	38.63	480.69
0.05	1.1893	21.28	180.53
0.025	1.0835	17.56	145.20
0.01	1.0313	15.92	131.43

Applying Theorem 1 to the sequence $\{Z_j\}_{j=1,\dots,L-1}$ defined in (4), from (5) we obtain an approximation and the associated error bound for $\mathbb{P}(S(m, T) \leq x)$ in the following way. Put for $s \in \{2, 3\}$,

$$Q_s = Q_s(x) = \mathbb{P} \left(\bigcap_{j=1}^{s-1} \{Z_j \leq x\} \right) \quad (12)$$

and observe that using the notation of (6) we have $Q_s = q_{s-1}$. For x such that $Q_2(x) \geq 1 - \alpha \geq 0.9$ we apply the result from Theorem 1 to obtain the approximation

$$\mathbb{P}(S(m, T) \leq x) \approx \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{(L-1)}}, \quad (13)$$

with an error bound of about $(L-1)F(\alpha_1, L-1)(1-Q_2)^2$. Observe that Q_2 and Q_3 represent the distributions of the scan statistics over sequences of variable Z_j of lengths $2m+k-3$ and respectively $3m+2k-5$. Q_2 and Q_3 are generally estimated by Monte Carlo simulation.

Thus, if $\hat{Q}_s$ is an estimate of Q_s , $s \in \{2, 3\}$, with $|\hat{Q}_s - Q_s| \leq \beta_s$ and x is such that $1 - \hat{Q}_2(x) \leq 0.1$ then

$$\left| \mathbb{P}(S(m, T) \leq x) - (2\hat{Q}_2 - \hat{Q}_3) [1 + \hat{Q}_2 - \hat{Q}_3 + 2(\hat{Q}_2 - \hat{Q}_3)^2]^{1-L} \right| \leq E_{total}, \quad (14)$$

where E_{total} is the total error of the approximation given by

$$E_{total} = (L-1) [\beta_2 + \beta_3 + F(\hat{Q}_2, L-1)(1 - \hat{Q}_2 + \beta_2)^2]. \quad (15)$$

One of the main advantage of this approximation method with respect to the product-type approximation proposed in Reference [12], who uses the same quantities Q_2 and Q_3 , is that it provides sharp error bounds for the approximation.

3. Some Related Problems to the Scan Statistics under Block-Factor Dependence Model

In order to illustrate the efficiency of the approximation (14) and the obtained error bounds, in this section we present two examples of statistics related to discrete scan statistics.

3.1. Length of the Longest Increasing Run in a i.i.d Sequence

Let $Y_1, Y_2, \dots, Y_{\tilde{T}}$ be a sequence of length $\tilde{T}$, $\tilde{T} \geq 1$, of independent and identically distributed random variables with the common distribution F . We say that the subsequence $(Y_k, \dots, Y_{k+l-1})$ forms an *increasing run* (or ascending run) of length $l \geq 1$, starting at position $k \geq 1$, if it verifies the following relation

$$Y_{k-1} > Y_k < Y_{k+1} < \dots < Y_{k+l-1} > Y_{k+l}. \quad (16)$$

We denote the length of the longest increasing run among the first $\tilde{T}$ random variables by $M_{\tilde{T}}$. This run statistics plays an important role in many applications in fields such computer science, reliability theory or quality control. The asymptotic behaviour of $M_{\tilde{T}}$ has been investigated by several authors depending on the common distribution, F . In the case of a continuous distribution [27] (see also Reference [28]) has shown that this behaviour does not depend on the common law. For the particular setting of uniform $U([0, 1])$ random variables, this problem was addressed by References [29–31]. Under the assumption that the distribution F is discrete, the limit behaviour of $M_{\tilde{T}}$ depends strongly on the common law F , as in Reference [32] (see also References [33,34]) proved for the case of geometric and Poisson distribution. In Reference [35], the case of discrete uniform distribution is investigated, while in Reference [36], the authors study the asymptotic distribution of $M_{\tilde{T}}$ when the variables are uniformly distributed but not independent.

In this section, we evaluate the distribution of the length of the longest increasing run using the methodology developed in Section 2. The idea is to express the distribution of the random variable $M_{\tilde{T}}$ in terms of the distribution of the scan statistics random variable.

Let $T = \tilde{T} - 1$ and define the block-factor transformation $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$f(x, y) = \begin{cases} 1, & \text{if } x < y \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

Then, our block-factor model becomes

$$X_i = f(Y_i, Y_{i+1}) = \mathbf{1}_{Y_i < Y_{i+1}}. \quad (18)$$

and $X_1, \dots, X_T$ form a 1-dependent and stationary sequence of random variables.

Notice that the distribution of $M_{\tilde{T}}$ and the distribution of the length of the longest run of ones, L_T , among the first T binary random variables X_i , are related and satisfy the following identity

$$\mathbb{P}(M_{\tilde{T}} \leq m) = \mathbb{P}(L_T < m), \text{ for } m \geq 1. \quad (19)$$

The statistics L_T is also known as the length of the longest success run or head run and was extensively studied in the literature. One can consult the monographs of References [16,17] for applications and further results concerning this statistic. Moreover, the random variable L_T can be interpreted as a particular case of the scan statistics random variable and between the two we have the relation

$$\mathbb{P}(L_T \geq m) = \mathbb{P}(S(m, T) = m). \quad (20)$$

Hence, combining (19) and (20), we can express the distribution of the length of the longest increasing run as

$$\mathbb{P}(M_{\tilde{T}} \leq m) = \mathbb{P}(S(m, T) < m). \quad (21)$$

Thus, we can estimate the distribution of $M_{\tilde{T}}$ using the foregoing identity and the approximations developed in Section 2 for the discrete scan statistics random variable.

We should also note that in Reference [30] the authors studied the asymptotic behaviour of L_T over a sequence of m -dependent binary random variables. They showed that, given a stationary m -dependent sequence of random variables with values 0 and 1, $\{V_k\}_{k \geq 1}$, if there exist positive constants t, C such that

$$\mathbb{P}(V_{k+1} = 1 | V_1 = \dots = V_k = 1) \geq \frac{1}{Ck^t}, \text{ for all } k \geq C, \quad (22)$$

then, as $N \rightarrow \infty$

$$\max_{1 \leq k \leq N} \left| \mathbb{P}(L_N < k) - e^{-Nr(k)} \right| = \mathcal{O} \left(\frac{(\ln N)^h}{N} \right), \quad (23)$$

where $r(k) = \mathbb{P}(V_1 = \dots = V_k = 1) - \mathbb{P}(V_1 = \dots = V_{k+1} = 1)$ and $h = \sup\{mt, 1\}$.

In order to illustrate the accuracy of the approximation of $M_{\tilde{T}}$ based on scan statistics, using the methodology developed in Section 2, we consider that the random variables Y_i 's have a common uniform $U([0, 1])$ distribution. Simple calculations show that $\mathbb{P}(X_1 = \dots = X_k = 1) = \frac{1}{(k+1)!}$ and

$$\mathbb{P}(X_{k+1} = 1 | X_1 = \dots = X_k = 1) = \frac{1}{k+2} \geq \frac{1}{2k}, \quad (24)$$

thus $C = 2$, $t = 1$ and $h = 1$. In the context of our particular situation, the result of Reference [30] in Equation (23) becomes:

$$\max_{1 \leq m \leq T} \left| \mathbb{P}(L_T < m) - e^{-Tr(m)} \right| = \mathcal{O} \left(\frac{\ln T}{T} \right), \quad (25)$$

where $r(m) = \mathbb{P}(X_1 = \dots = X_m = 1) - \mathbb{P}(X_1 = \dots = X_{m+1} = 1) = \frac{m+1}{(m+2)!}$.

In Table 2, we consider a numerical comparison study between the simulated value (column *Sim*) obtained by Monte Carlo simulation with $ITER_{sim} = 10^4$ trials, the approximation based on scan statistics (column *App*) computed from the Equation (14) where $\hat{Q}_2$ and $\hat{Q}_3$ are computed with $ITER_{app} = 10^5$ trials and the limit distribution (column *LimApp*) of the distribution of the length of the longest increasing run, $\mathbb{P}(M_{\tilde{T}} \leq m)$, in a sequence of $\tilde{T} = 10001$ random variables distributed uniformly over $[0, 1]$. The results show that both our method and the asymptotic approximation in (25) are very accurate. It is worth mentioning that for our simulations we used an adapted version of the Importance Sampling procedure introduced in Reference [3], an efficient method that proved to perform very well for small p values (where naive Monte Carlo methods tend to break down) [37].

Table 2. The distribution of the length of the longest increasing run: $\tilde{T} = 10001$, $ITER_{sim} = 10^4$, $ITER_{app} = 10^5$.

m	Sim	<i>App</i> Equation (14)	E_{total} Equation (15)	<i>LimApp</i> Equation (25)
5	0.00000700	0.00000733	0.14860299	0.00000676
6	0.17567262	0.17937645	0.01089628	0.17620431
7	0.80257424	0.80362353	0.00110990	0.80215088
8	0.97548510	0.97566460	0.00011579	0.97550345
9	0.99749821	0.99751049	0.00001114	0.99749792
10	0.99977074	0.99977183	0.00000098	0.99977038
11	0.99998075	0.99998083	0.00000008	0.99998073
12	0.99999851	0.99999851	0.00000001	0.99999851
13	0.99999989	0.99999989	0.00000000	0.99999989
14	0.99999999	0.99999999	0.00000000	0.99999999
15	1.00000000	1.00000000	0.00000000	1.00000000

3.2. The Moving Average - Like Process of Order q Model

We consider the particular situation of the one dimensional discrete scan statistics defined over a sequence of random variables obtained as a linear block factor of a discrete Gaussian white noise. Because of the similarity with the definition of a classical moving average process, we call that block factor model a *moving average - like* process. It is worth mentioning that the distribution of the scan statistics in the context of a moving average process for normal data was studied in Reference [38] where the authors compared the product-type approximation developed in Reference [13] with the approximation of Reference [23]. In the block-factor model introduced in (3), let $q \geq 1$ be a positive integer and $Y_1, Y_2, \dots, Y_{\tilde{T}}$ be a sequence of independent and identically Gaussian distributed random variables with known mean μ and variance σ^2 .

Let $\mathbf{a} = (a_1, \dots, a_{q+1}) \in \mathbb{R}^{q+1}$ be a fixed non null vector and take $f: \mathbb{R}^{q+1} \rightarrow \mathbb{R}$, the (measurable) transformation that defines the block-factor model, to be equal with

$$f(y_1, \dots, y_{q+1}) = a_1 y_1 + a_2 y_2 + \dots + a_{q+1} y_{q+1}. \quad (26)$$

For $i \in \{1, \dots, T\}$, with $T = \tilde{T} - q$, our dependent model is defined by the relation

$$X_i = a_1 Y_i + a_2 Y_{i+1} + \dots + a_{q+1} Y_{i+q}. \quad (27)$$

The moving sums of size m , W_i , $1 \leq i \leq T - m + 1$ can be expressed as

$$W_i = \sum_{j=i}^{i+m-1} X_j = b_1 Y_i + b_2 Y_{i+1} + \dots + b_{m+q} Y_{i+m-1+q}, \quad (28)$$

where the coefficients $b_1, \dots, b_{m+q}$ are evaluated by

(a) For $m \geq q$,

$$b_k = \begin{cases} \sum_{j=1}^k a_j & , k \in \{1, \dots, q\} \\ \sum_{j=1}^{q+1} a_j & , k \in \{q+1, \dots, m\} \\ \sum_{j=k-m+1}^{q+1} a_j & , k \in \{m+1, \dots, m+q\} \end{cases} \quad (29)$$

(b) For $m < q$,

$$b_k = \begin{cases} \sum_{j=1}^k a_j & , k \in \{1, \dots, m\} \\ \sum_{j=k-m+1}^k a_j & , k \in \{m+1, \dots, q\} \\ \sum_{j=k-m+1}^{q+1} a_j & , k \in \{q+1, \dots, m+q\}. \end{cases} \quad (30)$$

Therefore, for each $i \in \{1, \dots, T-m+1\}$, the random variable W_i follows a normal distribution with mean $\mathbb{E}[W_i] = (b_1 + \dots + b_{m+q})\mu$ and variance $\text{Var}[W_i] = (b_1^2 + \dots + b_{m+q}^2)\sigma^2$. Moreover, a simple calculation shows that the covariance matrix $\Sigma = \{\text{Cov}[W_t, W_s]\}$ has the entries

$$\text{Cov}[W_t, W_s] = \begin{cases} \left(\sum_{j=1}^{m+q-|t-s|} b_j b_{|t-s|+j} \right) \sigma^2 & , |t-s| \leq m+q-1 \\ 0 & , \text{otherwise.} \end{cases} \quad (31)$$

Given the mean and the covariance matrix of the vector $(W_1, \dots, W_{T-m+1})$, one can use the importance sampling algorithm developed in Reference [3] (see also Reference [37]) or the one presented in Reference [39] to estimate the distribution of the one dimensional discrete scan statistics $S(m, T)$. Another way is to use the quasi-Monte Carlo algorithm developed in Reference [40] to approximate the multivariate normal distribution.

In our application example we adopt the importance sampling procedure developed in Reference [3]. In order to evaluate the accuracy of the approximation developed in (14), we consider $q = 2$, $T = 1000$, $m = 20$, $Y_i \sim \mathcal{N}(0, 1)$ and the coefficients of the moving average model to be $(a_1, a_2, a_3) = (0.3, 0.1, 0.5)$. We compare our approximation (column *App*) given in (14) with the one (column *AppPT*) given in Reference [41] using product-type approximations. In Table 3, we present numerical results for the setting described above. In our algorithms we used $ITER_{app} = 10^6$ trials for the computation of $\hat{Q}_2(x)$ and $\hat{Q}_3(x)$ and $ITER_{sim} = 10^5$ trials for the Monte-Carlo simulation of the $\mathbb{P}(S(m, T) \leq x)$.

Table 3. MA-like($q = 2$) model: $m = 20$, $T = 1000$, $X_i = 0.3Y_i + 0.1Y_{i+1} + 0.5Y_{i+2}$, $ITER_{app} = 10^6$, $ITER_{sim} = 10^5$.

x	Sim	$AppPT$	App Equation (14)	E_{total} Equation (15)
11	0.582252	0.589479	0.584355	0.015156
12	0.770971	0.773700	0.771446	0.004010
13	0.889986	0.890009	0.889431	0.001167
14	0.951529	0.954536	0.951723	0.000370
15	0.980653	0.982433	0.980675	0.000124
16	0.992827	0.993690	0.992791	0.000042
17	0.997486	0.995471	0.997499	0.000014
18	0.999186	0.999411	0.999188	0.000004
19	0.999754	0.999717	0.999754	0.000001
20	0.999930	1	0.999930	0.000000

4. Conclusions

Block factor models defined from i.i.d. sequence generate random sequences with a particular type of dependence structure. For this type of dependence, the scan statistics can be viewed as the maximum of a 1-dependent stationary sequence, for which the distribution can be approximated with high accuracy. The approximation error can be controlled by using efficient algorithms of simulation as for example the importance sampling approach proposed in Reference [3] (see also Reference [25]). We approximated the distribution of longest increasing run statistics over an i.i.d sequence as a particular case of scan statistics distribution over a block factor model.

Author Contributions: Conceptualization, A.A. and C.P.; methodology, A.A. and C.P.; software, A.A.; validation, A.A. and C.P.; formal analysis, A.A. and C.P.; writing original draft preparation, C.P.; writing review and editing, A.A. and C.P.; visualization, A.A. and C.P.; supervision, A.A. and C.P.; project administration, C.P.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by a grant of the Romanian National Authority for Scientific Research and Innovation, project number POC P-37-257 and MCI National Core Program, project 25 N/2019 BIODIVERS 19270103.

Acknowledgments: The authors wish to thank the anonymous reviewers for their careful reading of the manuscript and their helpful suggestions and comments which led to the improvement of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Hoh, J.; Ott, J. Scan statistics to scan markers for susceptibility genes. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 9615–9617. [[CrossRef](#)] [[PubMed](#)]
- Sheng, K.-N.; Naus, J. Pattern matching between two non aligned random sequences. *Bull. Math. Biol.* **1994**, *56*, 1143–1162 [[CrossRef](#)]
- Naiman, D.Q.; Priebe, C.E. Computing scan statistic p values using importance sampling, with applications to genetics and medical image analysis. *J. Comput. Graph. Stat.* **2001**, *10*, 296–328. [[CrossRef](#)]
- Guerriero, M.; Pozdnyakov, V.; Glaz, J.; Willett, P. A repeated significance test with applications to sequential detection in sensor networks. *IEEE Trans. Signal Process.* **2010**, *58*, 3426–3435. [[CrossRef](#)]
- Guerriero, M.; Willett, P.; Glaz, J. Distributed target detection in sensor networks using scan statistics. *IEEE Trans. Signal Process.* **2009**, *57*, 2629–2639. [[CrossRef](#)]
- Darling, R.W.R.; Waterman, M.S. Extreme value distribution for the largest cube in a random lattice. *SIAM J. Appl. Math.* **1986**, *46*, 118–132. [[CrossRef](#)]
- Marcos, R.; Marcos, C. From star complexes to the field: open cluster families. *Astrophys. J.* **2008**, *672*, 342–351. [[CrossRef](#)]
- Boutsikas, M.V.; Koutras, M.V. Reliability approximation for Markov chain imbeddable systems. *Methodol. Comput. Appl. Probab.* **2000**, *2*, 393–411. [[CrossRef](#)]

9. Glaz, J.; Naus, J. Tight bounds and approximations for scan statistic probabilities for discrete data. *Ann. Appl. Probab.* **1991**, *1*, 306–318. [\[CrossRef\]](#)
10. Glaz, J.; Naus, J.; Wallenstein, S. *Scan Statistics*; Springer Series in Statistics; Springer: New York, NY, USA, 2001.
11. Glaz, J.; Balakrishnan, N. *Scan Statistics and Applications*; Springer Sciences+Business Media: Berlin, Germany, 1999.
12. Naus, J. Approximations for distributions of scan statistics. *J. Am. Stat. Assoc.* **1982**, *77*, 177–183. [\[CrossRef\]](#)
13. Wang, X.; Glaz, J.; Naus, J. Approximations and inequalities for moving sums. *Methodol. Comput. Appl. Probab.* **2012**, *14*, 597–616.
14. Chen, J.; Glaz, J. Scan statistics for monitoring data modeled by a negative binomial distribution. *Commun. Stat. Theory Methods* **2016**, *45*, 1632–1642. [\[CrossRef\]](#)
15. Naus, J. Probabilities for a generalized birthday problem. *J. Am. Stat. Assoc.* **1974**, *69*, 810–815. [\[CrossRef\]](#)
16. Balakrishnan, N.; Koutras, M.V. *Runs and Scans with Applications*; Wiley Series in Probability and Statistics; Wiley-Interscience [John Wiley & Sons]: New York, NY, USA, 2002.
17. Fu, J.C.; Lou, W. *Distribution Theory of Runs and Patterns and Its Applications: A Finite Markov Chain Imbedding Approach*; World Scientific Publishing Co., Inc.: River Edge, NJ, USA, 2003.
18. Ebneshrashoob, M.; Gao, T.; Wu, M. An efficient algorithm for exact distribution of discrete scan statistics. *Methodol. Comput. Appl. Probab.* **2005**, *7*, 1423–1436. [\[CrossRef\]](#)
19. Uchida, M. On generating functions of waiting time problems for sequence patterns of discrete random variables. *Ann. Inst. Stat. Math.* **1998**, *50*, 650–671. [\[CrossRef\]](#)
20. Chen, J.; Glaz, J. Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials. *Adv. Theory Pract. Stat.* **1997**, *1*, 285–298.
21. Arratia, R.; Gordon, L.; Waterman, M.S. The Erdos-Rényi law in distribution for coin tossing and sequence matching. *Ann. Stat.* **1990**, *18*, 539–570. [\[CrossRef\]](#)
22. Burton, R.M.; Goulet, M.; Meester, R. On one-dependent processes and k-block factors. *Ann. Probab.* **1993**, *21*, 2157–2168. [\[CrossRef\]](#)
23. Haiman, G.; Preda, C. One dimensional scan statistics generated by some dependent stationary sequences. *Stat. Probab. Lett.* **2013**, *83*, 1457–1463. [\[CrossRef\]](#)
24. Amárioarei, A. Approximation for the Distribution of Extremes of One Dependent Stationary Sequences of Random Variables. *arXiv* **2012**, arXiv:1211.5456v1.
25. Amárioarei, A. Approximations for the Multidimensional Discrete Scan Statistics. Ph.D. Thesis, University of Lille, Lille, France, 2014.
26. Haiman, G. Estimating the distributions of scan statistics with high precision. *Extremes* **2000**, *3*, 349–361. [\[CrossRef\]](#)
27. Pittel, B. Limiting behavior of a process of runs. *Ann. Probab.* **1981**, *9*, 119–129. [\[CrossRef\]](#)
28. Frolov, A.; Martikainen, A. On the length of the longest increasing run in R^d . *Stat. Prob. Lett.* **1999**, *41*, 153–161. [\[CrossRef\]](#)
29. Grill, K. Erdos-Révész type bounds for the length of the longest run from a stationary mixing sequence. *Probab. Theory Relat. Fields* **1987**, *75*, 169–179. [\[CrossRef\]](#)
30. Novak, S. Longest runs in a sequence of m -dependent random variables. *Probab. Theory Relat. Fields* **1992**, *91*, 269–281. [\[CrossRef\]](#)
31. Révész, P. Three problems on the length of increasing runs. *Stochastic Process. Appl.* **1983**, *5*, 169–179. [\[CrossRef\]](#)
32. Csaki, E.; Foldes, A. On the length of the longest monotone block. *Studia Scientiarum Mathematicarum Hungarica* **1996**, *31*, 35–46.
33. Eryilmaz, S. A note on runs of geometrically distributed random variables. *Discrete Math.* **2006**, *306*, 1765–1770. [\[CrossRef\]](#)
34. Grabner, P.; Knopfmacher, A.; Prodinger, H. Combinatorics of geometrically distributed random variables: Run statistics. *Theoret. Comput. Sci.* **2003**, *297*, 261–270. [\[CrossRef\]](#)
35. Louchard, G. Monotone runs of uniformly distributed integer random variables: A probabilistic analysis. *Theoret. Comput. Sci.* **2005**, *346*, 358–387. [\[CrossRef\]](#)
36. Mitton, N.; Paroux, K.; Sericola, B.; Tixeuil, S. Ascending runs in dependent uniformly distributed random variables: application to wireless networks. *Methodol. Comput. Appl. Probab.* **2010**, *12*, 51–62. [\[CrossRef\]](#)

37. Malley, J.; Naiman, D.Q.; Bailey-Wilson, J. A compresive method for genome scans. *Hum. Heredity* **2002**, *54*, 174–185. [[CrossRef](#)] [[PubMed](#)]
38. Wang, X.; Zhao, B.; Glaz, J. A multiple window scan statistic for time series models. *Stat. Probab. Lett.* **2014**, *94*, 196–203. [[CrossRef](#)]
39. Shi, J.; Siegmund, D.; Yakir, B. Importance sampling for estimating p values in linkage analysis. *J. Am. Stat. Assoc.* **2007**, *102*, 929–937. [[CrossRef](#)]
40. Genz, A.; Bretz, F. *Computation of Multivariate Normal and T Probabilities*; Springer: New York, NY, USA, 2009.
41. Wang, X.; Glaz, J. A variable window scan statistic for $MA(1)$ process. In Proceedings of the 15th International Conference on Applied Stochastic Models and Data Analysis ASMDA 2013, Barcelona, Spain, 25–28 June 2013; pp. 955–962.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).