Approximations for two-dimensional discrete scan statistics in some dependent models

Alexandru Amărioarei[1],[2] and Cristian Preda[1]

[1] Laboratoire de Mathématiques Paul Painlevé, UMR CNRS 8524, MODAL-INRIA Lille, EA 2694 Université des Sciences et Technologies de Lille, France
[2] National Institute of R&D for Biological Sciences, Bucharest, Romania
E-mail: alexandru.amarioarei@inria.fr
E-mail: cristian.preda@polytech-lille.fr

Abstract. We consider the two-dimensional discrete scan statistic generated by block factors from i.i.d. sequences. We present the approximation for the distribution of the scan statistics and error bounds. A simulation study illustrates our methodology.
Keywords: Scan statistics, m-dependent sequences, block-factor.

1 Introduction

Let N_1, N_2 be positive integers and $\{X_{i,j} \mid 1 \leq i \leq N_1, 1 \leq j \leq N_2\}$ be a family of nonnegative integer random variables from a specified distribution. For $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$, X_{ij} represents the number of some events observed in the elementary square sub-region $[i, i+1] \times [j, j+1]$. Let m_1, m_2 be positive integers, $1 \leq m_1 \leq N_1$, $1 \leq m_2 \leq N_2$. For $1 \leq t \leq N_1 - m_1 + 1$, $1 \leq s \leq N_2 - m_2 + 1$ let

$$\nu_{ts} = \nu_{t,s}(m_1, m_2) = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} X_{ij}. \tag{1}$$

The *two-dimensional discrete scan statistic* is defined as the largest number of events in any $m_1 \times m_2$ rectangular scanning window within the rectangular region $[1, N_1] \times [1, N_2]$, i.e.

$$S = S(m_1, m_2, N_1, N_2) = \max_{\substack{1 \leq t \leq N_1 - m_1 + 1 \\ 1 \leq s \leq N_2 - m_2 + 1}} \nu_{ts}. \tag{2}$$

Most of research devoted to two-dimensional discrete scan statistics considers the i.i.d. model for $X_{i,j}$'s. Then, the statistic S is used for testing the null hypothesis that the $X_{i,j}$'s are independent and identically distributed according to some specified probability law, in general Bernoulli, binomial or Poisson (see Glaz et al.[4]). Since there are no exact formulas for $P(S \leq n)$, various methods of approximation and bounds for $P(S \leq n)$ have been proposed by several authors. An overview of these methods as well as a complete bibliography on the subject are given in Chen and Glaz [3], Glaz et al. [4], Boutsikas and Koutras [2], Haiman and Preda [7] and references therein. In this paper

we consider the two-dimensional discrete scan statistics generated by some particular dependent sequences $\{X_{i,j} \mid 1 \leq i \leq N_1, 1 \leq j \leq N_2\}$. More precisely, we are interested in the case where the $X_{i,j}$'s are obtained as block-factors from an independent and identically distributed sequence of random variables $\{Y_{i,j} \mid 0 \leq i \leq N_1+1, 0 \leq j \leq N_2+1\}$, in the following way:

$$X_{i,j} = f(Y_{i,j}, Y_{i,j-1}, Y_{i,j+1}, Y_{i-1,j-1}, Y_{i-1,j}, Y_{i-1,j+1}, Y_{i+1,j-1}, Y_{i+1,j}, Y_{i+1,j+1}), \tag{3}$$

where f is a non-negative measurable function $f : \mathbb{R}^9 \to \mathbb{R}_+$ (see Figure 1).

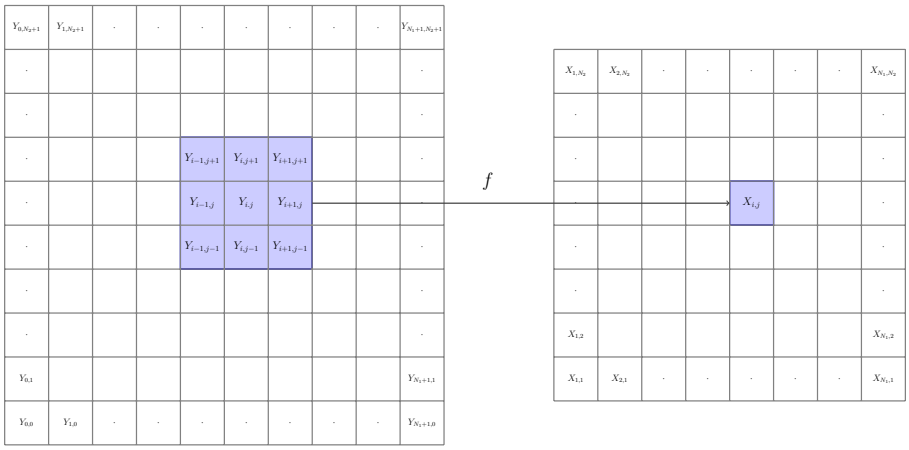

Fig. 1. Two-dimensional block-factor model

The paper is organized as follows. In Section 2, we present the methodology for approximating the distribution of the scan statistics generated by the model (3) providing approximations and error bounds. Numerical results based on simulations are presented in Section 3 for particular block-factor functions, f.

2 The scan statistics as an extremum of a 1-dependent stationary sequence of random variables

The methodology we use to approximate the distribution of scan statistics generated by the dependent model (3) follows closely to that presented in Haiman and Preda [7] for the i.i.d. model. Let us suppose that $N_1 = (L_1+1)(m_1+1)-2$, $N_2 = (L_2+1)(m_2+1)-2$ where L_1 and L_2 are positive integers and define

$$Z_k = \max_{\substack{1 \leq t \leq L_1(m_1+1) \\ (k-1)(m_2+1)+1 \leq s \leq k(m_2+1)}} \nu_{ts}, \quad k \in \{1, 2, \ldots, L_2\}. \tag{4}$$

The random variables Z_k represent the scan statistics on the overlapping $N_1 \times 2m_2$ rectangular regions

$$\mathcal{R}_k = [1, N_1] \times [(k-1)(m_2+1)+1, (k+1)(m_2+1)-2].$$

From the independence of Y_{ij} and

$$\begin{aligned} Z_{k-1} &\in \sigma\left(X_{ij} \mid 1 \le i \le N_1,\ (k-2)(m_2+1)+1 \le j \le k(m_2+1)-2\right) \\ &\in \sigma\left(Y_{ij} \mid 0 \le i \le N_1+1,\ (k-2)(m_2+1) \le j \le k(m_2+1)-1\right), \end{aligned} \tag{5}$$

$$\begin{aligned} Z_{k} &\in \sigma\left(X_{ij} \mid 1 \le i \le N_1,\ (k-1)(m_2+1)+1 \le j \le (k+1)(m_2+1)-2\right) \\ &\in \sigma\left(Y_{ij} \mid 0 \le i \le N_1+1,\ (k-1)(m_2+1) \le j \le (k+1)(m_2+1)-1\right), \end{aligned} \tag{6}$$

$$\begin{aligned} Z_{k+1} &\in \sigma\left(X_{ij} \mid 1 \le i \le N_1,\ k(m_2+1)+1 \le j \le (k+2)(m_2+1)-2\right) \\ &\in \sigma\left(Y_{ij} \mid 0 \le i \le N_1+1,\ k(m_2+1) \le j \le (k+2)(m_2+1)-1\right), \end{aligned} \tag{7}$$

we can verify that the sequence $(Z_k)_{1\le k\le L_2}$ is 1-dependent (see Figure 2 for the case $k=2$). Since Y_{ij} are also identically distributed we conclude that the sequence Z_k is stationary.

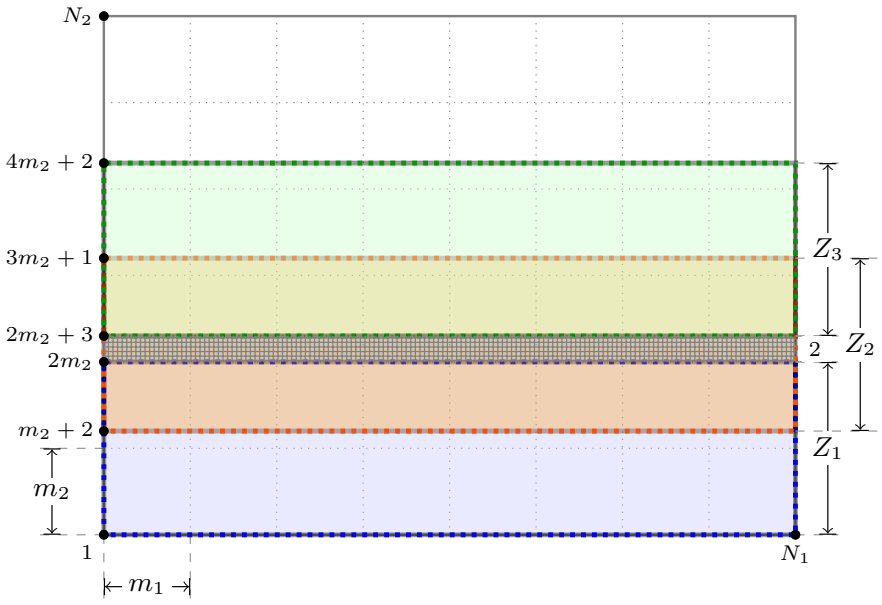

Fig. 2. Illustration of Z_k emphasizing the 1-dependence

Notice that from (2), (4) and the measurability of f, the following relation holds

$$S = \max_{1\le k\le L_2} Z_k. \tag{8}$$

The relation described in (8) is the key idea behind our approximation. The methodology is based on the following result developed in Haiman [5, Theorem 4] and improved in Amarioarei [1, Theorem 2.6]:
Let $(T_k)_{k\ge 1}$ be a strictly stationary 1-dependent sequence of random variables and for $x < \sup\{u|\mathbb{P}(T_1 \le u) < 1\}$, consider

$$q_m = q_m(x) = \mathbb{P}(\max(T_1, \ldots, T_m) \le x). \tag{9}$$

Theorem 1. *Assume that x is such that $q_1(x) \geq 1-\alpha \geq 0.9$ and define $\eta = 1 + l\alpha$ with $l = l(\alpha) > t_2^3(\alpha)$ and $t_2(\alpha)$ the second root in magnitude of the equation $\alpha t^3 - t + 1 = 0$. Then the following relation holds*

$$\left| q_m - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^m} \right| \leq mF(\alpha, m)(1 - q_1)^2, \tag{10}$$

with

$$F(\alpha, m) = 1 + \frac{3}{m} + \left[\frac{\Gamma(\alpha)}{m} + K(\alpha) \right] (1 - q_1), \tag{11}$$

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

$$K(\alpha) = \frac{\frac{11-3\alpha}{(1-\alpha)^2} + 2l(1+3\alpha)\frac{2+3l\alpha-\alpha(2-l\alpha)(1+l\alpha)^2}{[1-\alpha(1+l\alpha)^2]^3}}{1 - \frac{2\alpha(1+l\alpha)}{[1-\alpha(1+l\alpha)^2]^2}}, \tag{12}$$

$$\begin{aligned} L(\alpha) &= 3K(\alpha)(1 + \alpha + 3\alpha^2)[1 + \alpha + 3\alpha^2 + K(\alpha)\alpha^3] + \alpha^6 K^3(\alpha), \\ &\quad + 9\alpha(4 + 3\alpha + 3\alpha^2) + 55.1 \end{aligned} \tag{13}$$

$$E(\alpha) = \frac{\eta^5 \left[1 + (1-2\alpha)\eta\right]^4 \left[1 + \alpha(\eta - 2)\right] \left[1 + \eta + (1-3\alpha)\eta^2\right]}{2(1 - \alpha\eta^2)^4 \left[(1 - \alpha\eta^2)^2 - \alpha\eta^2(1 + \eta - 2\alpha\eta)^2\right]}. \tag{14}$$

Let define for $r \in \{2, 3\}$,

$$Q_r = Q_r(n) = \mathbb{P}\left(\bigcap_{k=1}^{r-1} \{Z_k \leq n\} \right) = \mathbb{P}\left(\max_{\substack{1 \leq t \leq L_1(m_1+1) \\ 1 \leq s \leq (r-1)(m_2+1)}} \nu_{ts} \leq n \right). \tag{15}$$

For n such that $Q_2(n) \geq 1 - \alpha_1 \geq 0.9$, we can apply the result in Theorem 1 to obtain the first step approximation

$$\left| \mathbb{P}(S \leq n) - \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{L_2}} \right| \leq L_2 F(\alpha_1, L_2)(1 - Q_2)^2. \tag{16}$$

In order to evaluate the approximation in (16) one has to find approximations for the quantities Q_2 and Q_3. To achieve this, we will apply for the second time the result in Theorem 1. We define, as in (4), for each $r \in \{2, 3\}$ and $l \in \{1, 2, \ldots, L_1\}$ the sequences

$$Z_l^{(r)} = \max_{\substack{(l-1)(m_1+1)+1 \leq t \leq l(m_1+1) \\ 1 \leq s \leq (r-1)(m_2+1)}} \nu_{ts}. \tag{17}$$

As for the sequence Z_k, we deduce that the sequences $Z_l^{(r)}$ in (17) are stationary, 1-dependent and the following relation holds:

$$Q_r = \mathbb{P}\left(\max_{1 \leq l \leq L_1} Z_l^{(r)} \leq n \right), \quad r \in \{2, 3\}. \tag{18}$$

Denoting, for $u, r \in \{2,3\}$

$$Q_{ur} = Q_{ur}(n) = \mathbb{P}\left(\bigcap_{l=1}^{u-1}\{Z_l^{(r)} \le n\}\right) = \mathbb{P}\left(\max_{\substack{1\le t\le (u-1)(m_1+1)\\ 1\le s\le (r-1)(m_2+1)}} \nu_{ts} \le n\right) \quad (19)$$

then, under the supplementary condition that $Q_{23}(n) \ge 1-\alpha_2 \ge 0.9$, we apply Theorem 1 to obtain

$$\left| Q_r - \frac{2Q_{2r} - Q_{3r}}{\left[1 + Q_{2r} - Q_{3r} + 2(Q_{2r} - Q_{3r})^2\right]^{L_1}} \right| \le L_1 F(\alpha_2, L_1)(1 - Q_{2r})^2. \quad (20)$$

Combining (16) and (20) we find an approximation formula for the distribution of the scan statistic depending on the values of $\{Q_{22}, Q_{23}, Q_{32}, Q_{33}\}$. There are no exact formulas for Q_{ur}, $u, r \in \{2,3\}$, thus they will be evaluated by Monte Carlo simulation.

For the error computation we have to notice that there are three expressions involved: the first one is the *theoretical error* (E_{app}) obtained from the substitution of (20) in (16) and the other two are *simulations errors*, one corresponding to the approximation formula (E_{sf}) and the other to the error formula (E_{sapp}). In what follows we will deal with each of them separately. To simplify the presentation it will be convenient to introduce the following notations;

$$A(x,y,m) = \frac{2x-y}{[1+x-y+2(x-y)^2]^m},\ \alpha_1 = 1 - Q_3,\ \alpha_2 = 1 - Q_{23},$$
$$F_1 = F(\alpha_2, L_1),\ F_2 = F(\alpha_1, L_2),\ R_s = A\left(Q_{2s}, Q_{3s}, L_1\right),\ s \in \{2,3\}.$$

It is not hard to see (based on mean value theorem in two dimensions) that if $y_i \le x_i$, $i \in \{1,2\}$, then we have the inequality:

$$|A(x_1,y_1,m) - A(x_2,y_2,m)| \le m\left[|x_1 - x_2| + |y_1 - y_2|\right]. \quad (21)$$

From (16) and (21) we get

$$\begin{aligned} |\mathbb{P}(S \le n) - A\left(R_2, R_3, L_2\right)| &\le |\mathbb{P}(S \le n) - A\left(Q_2, Q_3, L_2\right)| + \\ &\quad |A\left(Q_2, Q_3, L_2\right) - A\left(R_2, R_3, L_2\right)| \\ &\le L_2 F_2 \left(1 - Q_2\right)^2 + L_2\left[|Q_2 - R_2| + |Q_3 - R_3|\right] \end{aligned} \quad (22)$$

If we substitute (20) in (22) and we take $B_2 = 1 - R_2 + L_1 F_1 (1 - Q_{22})^2$, then the theoretical approximation error is given by

$$E_{app} = L_2 F_2 B_2^2 + L_1 L_2 F_1\left[(1 - Q_{22})^2 + (1 - Q_{23})^2\right]. \quad (23)$$

To compute the simulation error corresponding to the approximation formula let's denote by $\hat{Q}_{ur}$ the simulated values of Q_{ur} for each $u, r \in \{2,3\}$. Usually between the true and estimated values we have

$$\left| Q_{ur} - \hat{Q}_{ur} \right| \le \beta_{ur}. \quad (24)$$

Indeed, if $ITER$ is the number of iterations used in the Monte Carlo simulation algorithm for estimation of Q_{ur} then, one can consider, for example, the bound $\beta_{ur} = 1.96\sqrt{\frac{\hat{Q}_{ur}(1-\hat{Q}_{ur})}{ITER}}$ with a 95% confidence level. Taking for $r \in \{2,3\}$, $\hat{Q}_r = A\left(\hat{Q}_{2r}, \hat{Q}_{3r}, L_1\right)$ to be the simulated values that corresponds to Q_r and applying (19) whenever $\hat{Q}_3 \leq \hat{Q}_2$ we have

$$\begin{aligned}\left|A\left(R_2,R_3,L_2\right)-A\left(\hat{Q}_2,\hat{Q}_3,L_2\right)\right| &\leq L_2\left[\left|R_2-\hat{Q}_2\right|+\left|R_3-\hat{Q}_3\right|\right]\\ &\leq L_1L_2\left[\left|Q_{22}-\hat{Q}_{22}\right|+\left|Q_{23}-\hat{Q}_{23}\right|+\right.\\ &\left.\left|Q_{32}-\hat{Q}_{32}\right|+\left|Q_{33}-\hat{Q}_{33}\right|\right]. \qquad (25)\end{aligned}$$

From (25) and (24) we obtain the first simulation error

$$E_{sf} = L_1L_2(\beta_{22}+\beta_{23}+\beta_{32}+\beta_{33}). \qquad (26)$$

Finally, introducing

$$\begin{aligned}C_{2r} &= 1-\hat{Q}_{2r}+\beta_{2r}, \quad r\in\{2,3\},\\ C_2 &= 1-\hat{Q}_2+L_1(\beta_{22}+\beta_{32})+L_1F_1C_{22}^2,\end{aligned}$$

and substituting them in (23), we get the simulation error associated with the approximation error formula

$$E_{sapp} = L_2F_2C_2^2 + L_1L_2F_1\left[C_{22}^2+C_{23}^2\right]. \qquad (27)$$

Adding the expressions from (23), (26) and (27) we obtain the total error,

$$E_{total} = E_{app}+E_{sf}+E_{sapp}. \qquad (28)$$

3 Simulation study

In order to illustrate the results presented in Section 2, we consider the following block-factor dependence model. Let $\{Y_{ij} \mid i=0,\ldots,N_1+1,\ j=0,\ldots,N_2+1\}$ be an i.i.d. sequence of Bernoulli r.v.'s with parameter p. For each $1\leq i\leq N_1$ and $1\leq j\leq N_2$, define the r.v.'s X_{ij} by

$$X_{ij} = \begin{cases} 1, \text{ if } Y_{ij}=1 \text{ and } \displaystyle\sum_{k\in\{-1,0,1\}} Y_{i+k,j+k} \geq 2,\\ 0, \text{ otherwise.}\end{cases} \qquad (29)$$

The model in (29) is a particular case of (3). Obviously X_{ij}'s are dependent Bernoulli r.v.'s with parameter $p' = p\left[1-(1-p)^8\right]$. A success of X_{ij} occurs each time when $Y_{ij}=1$ and there is at least one success in its neighborhood (horizon one).

In our setting we consider the scanning window of size $m_1\times m_2 = 4\times 6$ and the region to be scanned of size 53×75 ($L_1=L_2=10$). The numerical results

presented in Table 1 corresponds to $p \in \{0.01, 0.1\}$. The second column in Table 1 (*Sim Dep*) corresponds to the estimate of $\mathbb{P}(S \leq n)$ with 10^5 trials. The column *Approx Dep* presents the approximations obtained by our methodology. Columns $4-6$ are the associated errors computed in Section 2 with $ITER = 10^9$. We compare the dependent model in (29) with the Bernoulli independent model with parameter p' using the results in Haiman and Preda [7]. The results are presented in the last two columns of Table 1. Figure 3 presents the ecdf of the scan statistic computed by simulation (*Sim Dep* and *Sim Indep*).

n	*Sim Dep*	*Approx Dep*	E_{app} (23)	E_{sim} (26)+(27)	E_{total} (28)	*Sim Indep*	*Approx Indep*
		$p = 0.01,\ p' = 0.00077$					
2	0.91937	0.91959	0.00351	0.00167	0.00518	0.99956	0.99921
3	0.98750	0.98748	0.00004	0.00046	0.00051	1	0.99999
4	0.99930	0.99915	0.00000	0.00010	0.00010	1	1
5	0.99993	0.99993	0.00000	0.00002	0.00002	1	1
		$p = 0.1,\ p' = 0.05695$					
9	0.93423	0.93247	0.00120	0.00111	0.00231	0.99957	0.99941
10	0.98847	0.98780	0.00003	0.00042	0.00045	0.99999	0.99995
11	0.99815	0.99812	0.00000	0.00015	0.00015	1	1
12	0.99971	0.99984	0.00000	0.00004	0.00004	1	1
13	0.99996	0.99999	0.00000	0.00001	0.00001	1	1

Table 1. Values for $\mathbb{P}(S \leq n)$ when $m1 = 4$, $m2 = 6$, $L1 = 10$, $L2 = 10$, $ITER = 10^9$

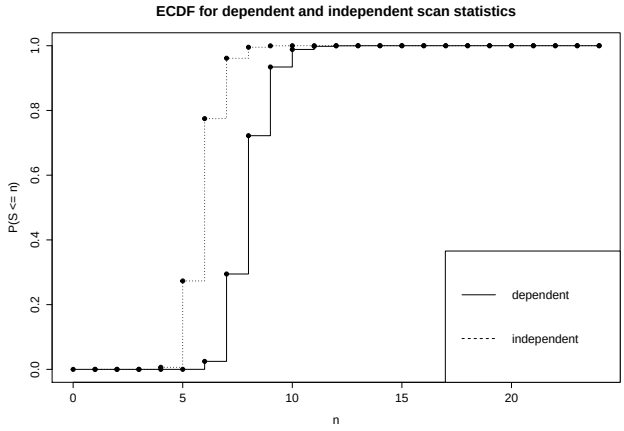

Fig. 3. Empirical cumulative distribution functions for $p = 0.1$.

References

1.Amarioarei, A.: *Approximation for the distribution of extremes of one dependent stationary sequences of random variables*, arXiv:1211.5456v1 (submitted)

2.Boutsikas, M. and Koutras, M. *Bounds for the distribution of two dimensional binary scan statistics*, Probability in the Engineering and Information Sciences, 17, 509–525, 2003.

3.Chen, J. and Glaz, J. *Two-dimensional discrete scan statistics*, Statistics and Probability Letters 31, 59–68, 1996.

4.Glaz, J., Naus, J., Wallenstein, S. *Scan Statistics*, Springer Series in Statistics, Springer-Verlag, New York, Inc., 2001.

5.Haiman, G. *First passage time for some stationary processes*, Stochastic Processes and their Applications, 80 (2), 231-248, 1999.

6.Haiman, G. *Estimating the distributions of scan statistics with high precision*, Extremes 3:4, 349-361, 2000.

7.Haiman, G. and Preda, C. *Estimation for the distribution of two dimensional discrete scan statistics*, Methodology and Computing in Applied Probability, Vol. 8, No. 3, 373-382, 2006.