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

Alexandru Amărioarei
Cristian Preda

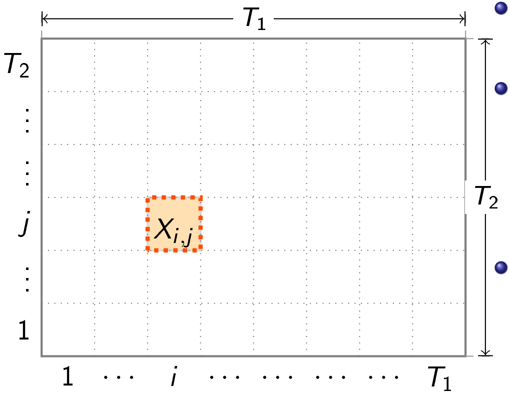
Laboratoire de Mathématiques Paul Painlevé
Département de Probabilités et Statistique
Université de Lille 1, INRIA Modal Team

16^{th} SPSR Conference
26 April, 2013, București, România

Outline

Outline

Introducing the General Model



Let T_1, T_2 be positive integers

- Rectangular region $\mathcal{R} = [0, T_1] \times [0, T_2]$
- $(X_{ij})_{\substack{1 \leq i \leq T_1 \\ 1 \leq j \leq T_2}}$ integer r.v.'s
  - Bernoulli$(\mathcal{B}(1, p))$
  - Binomial$(\mathcal{B}(n, p))$
  - Poisson$(\mathcal{P}(\lambda))$
- X_{ij} number of observed events in the elementary subregion $r_{ij} = [i - 1, i] \times [j - 1, j]$

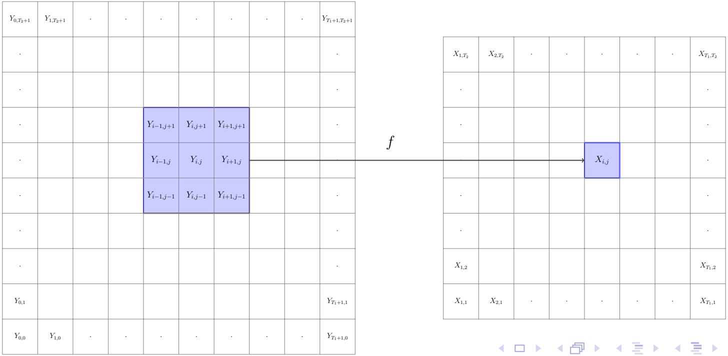
Introducing the Block-Factor Model

Consider for $1 \leq i \leq T_1, 1 \leq j \leq T_2$ the following block-factor model:

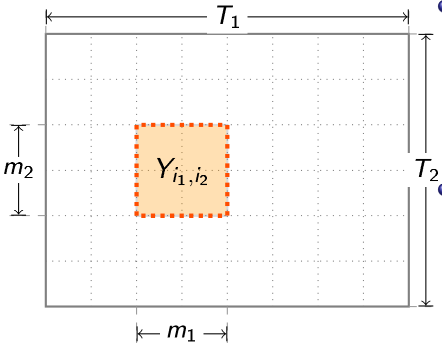
$$X_{i,j} = f(Y_{i,j}, Y_{i,j-1}, Y_{i,j+1}, Y_{i-1,j-1}, Y_{i-1,j}, Y_{i-1,j+1}, Y_{i+1,j-1}, Y_{i+1,j}, Y_{i+1,j+1}),$$

with $f : \mathbb{R}^9 \to \mathbb{R}_+$ and i.i.d. sequence

$$\{Y_{i,j} \mid 0 \leq i \leq T_1 + 1, 0 \leq j \leq T_2 + 1\}$$



Defining the Scan Statistic



Let m_1, m_2 be positive integers

- Define for $1 \leq i_j \leq T_j - m_j + 1$,

$$Y_{i_1 i_2} = \sum_{i=i_1}^{i_1+m_1-1} \sum_{j=i_2}^{i_2+m_2-1} X_{ij}$$

- The two dimensional scan statistic,

$$S_{m_1,m_2}(T_1, T_2) = \max_{\substack{1 \leq i_1 \leq T_1 - m_1 + 1 \\ 1 \leq i_2 \leq T_2 - m_2 + 1}} Y_{i_1 i_2}$$

- Used for testing the null hypotheses of randomness against the alternative hypothesis of clustering

Outline

Problem and related results

Problem

Approximate the distribution of two dimensional discrete scan statistic for the block-factor model

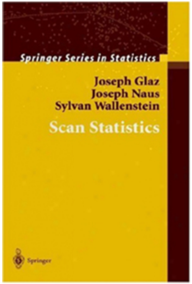
$$\mathbb{P}\left(S_{m_1,m_2}(T_1,T_2)\leq n\right).$$

- Dependent model: **no results !**
- Independent model:
  - No exact formulas
  - For Bernoulli case:
    - product type approximations (Boutsikas and Koutras 2000)
    - Poisson approximations (Chen and Glaz 1996)
    - bounds (Boutsikas and Koutras 2003)
  - For binomial and Poisson cases: (Glaz 2009)
    - Product type approximation
    - Lower bound

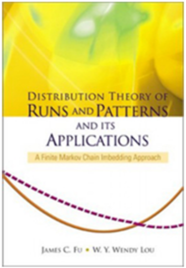
Literature


STATISTICS FOR INDUSTRY AND TECHNOLOGY
Joseph Glaz
N. Balakrishnan
Editors
Scan Statistics and Applications
Birkhäuser


Runs and Scans with Applications
N. Balakrishnan
Markos V. Koutras
WILEY SERIES IN PROBABILITY AND STATISTICS


Springer Series in Statistics
Joseph Glaz
Joseph Naus
Sylvan Wallenstein
Scan Statistics
Springer


STATISTICS FOR INDUSTRY AND TECHNOLOGY
Joseph Glaz
Vladimir Pozdnyakov
Sylvan Wallenstein
Editors
Scan Statistics
Methods and Applications
Birkhäuser


DISTRIBUTION THEORY OF RUNS AND PATTERNS AND ITS APPLICATIONS
A Finite Markov Chain Imbedding Approach
JAMES C. FU • W. Y. WENDY LOU

Outline

Key Idea

Haiman(2000) proposed a different approach

Main Observation

The scan statistic r.v. can be viewed as a maximum of a sequence of 1-dependent stationary r.v..

- The idea:
  - discrete and continuous one dimensional scan statistic: Haiman (2000,2007)
  - discrete and continuous two dimensional scan statistic: Haiman and Preda (2002,2006)
  - discrete three dimensional scan statistic: Amarioarei (2013)

Writing the Scan as an Extreme of 1-Dependent R.V.'s

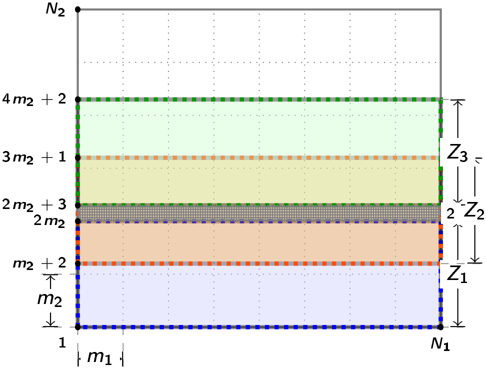
Let $T_j = (L_j + 1)(m_j + 1) - 2$, $j \in \{1, 2\}$ positive integers

- Define for $l \in \{1, 2, \dots, L_2\}$

$$Z_l = \max_{\substack{1 \le i_1 \le L_1(m_1+1) \\ (l-1)(m_2+1)+1 \le i_2 \le l(m_2+1)}} Y_{i_1 i_2}$$

- $(Z_l)_l$ is 1-dependent and stationary
- Observe

$$S_{m_1,m_2}(T_1, T_2) = \max_{1 \le l \le L_2} Z_l$$



Main Tool

Let $(Z_j)_{j\geq 1}$ be a strictly stationary 1-dependent sequence of r.v.'s and let $q_m = q_m(x) = \mathbb{P}(\max(Z_1, \ldots, Z_m) \leq x)$, with $x < \sup\{u | \mathbb{P}(Z_1 \leq u) < 1\}$.

Main Theorem (Haiman 1999, Amarioarei 2012)

For x such that $\mathbb{P}(Z_1 > x) = 1 - q_1 \leq \alpha < 0.1$ and $m > 3$ we have

$$\left| q_m - \frac{6(q_1 - q_2)^2 + 4q_3 - 3q_4}{(1 + q_1 - q_2 + q_3 - q_4 + 2q_1^2 + 3q_2^2 - 5q_1q_2)^m} \right| \leq \Delta_1(1 - q_1)^3,$$

$$\left| q_m - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^m} \right| \leq \Delta_2(1 - q_1)^2,$$

- $\Delta_1 = \Delta_1(\alpha, q_1, m) = \Gamma(\alpha) + mK(\alpha)$
- $\Delta_2 = mE(\alpha, q_1, m) = m\left[1 + \frac{3}{m} + K(\alpha)(1 - q_1) + \frac{\Gamma(\alpha)(1-q_1)}{m}\right]$.

Outline

First Step Approximation

Using Main Theorem we obtain

- Define
$$Q_2 = \mathbb{P}(Z_1 \leq k)$$
$$Q_3 = \mathbb{P}(Z_1 \leq k, Z_2 \leq k)$$

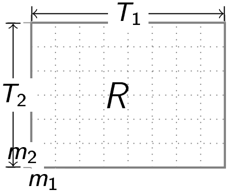
- If $1 - Q_2 \leq \alpha_1 < 0.1$ the (first) approximation

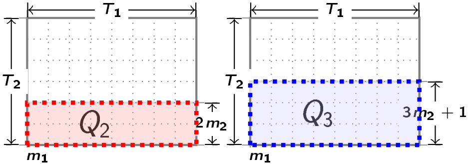
$$\mathbb{P}(S \leq k) \approx \frac{2Q_2 - Q_3}{[1 + Q_2 - Q_3 + 2(Q_2 - Q_3)^2]^{L_2}}$$

where $S = S_{m_1,m_2}(T_1, T_2)$

- Approximation error

$$L_2 E(\alpha_1, L_2)(1 - Q_2)^2$$





Second Step Approximation

Q_2 :

- For $s \in \{1, 2, \ldots, L_1\}$

$$Z_s^{(2)} = \max_{\substack{(s-1)(m_1+1)+1 \leq i_1 \leq s(m_1+1) \\ 1 \leq i_2 \leq m_2+1}} Y_{i_1 i_2}$$

- $$Q_2 = \mathbb{P}\left(\max_{1 \leq s \leq L_1} Z_s^{(2)} \leq k\right)$$

- Define
$$Q_{22} = \mathbb{P}(Z_1^{(2)} \leq k)$$
$$Q_{32} = \mathbb{P}(Z_1^{(2)} \leq k, Z_2^{(2)} \leq k)$$

- Approximation $(1 - Q_{22} \leq \alpha_2)$

$$Q_2 \approx \frac{2Q_{22} - Q_{32}}{\left[1 + Q_{22} - Q_{32} + 2(Q_{22} - Q_{32})^2\right]^{L_1}}$$

- Error

$$L_1 E(\alpha_2, L_1)(1 - Q_{22})^2$$

Q_3 :

- For $s \in \{1, 2, \ldots, L_1\}$

$$Z_s^{(3)} = \max_{\substack{(s-1)(m_1+1)+1 \leq i_1 \leq s(m_1+1) \\ 1 \leq i_2 \leq 2(m_2+1)}} Y_{i_1 i_2}$$

- $$Q_3 = \mathbb{P}\left(\max_{1 \leq l \leq L_1} Z_s^{(3)} \leq k\right)$$

- Define
$$Q_{23} = \mathbb{P}(Z_1^{(3)} \leq k)$$
$$Q_{33} = \mathbb{P}(Z_1^{(3)} \leq k, Z_2^{(3)} \leq k)$$

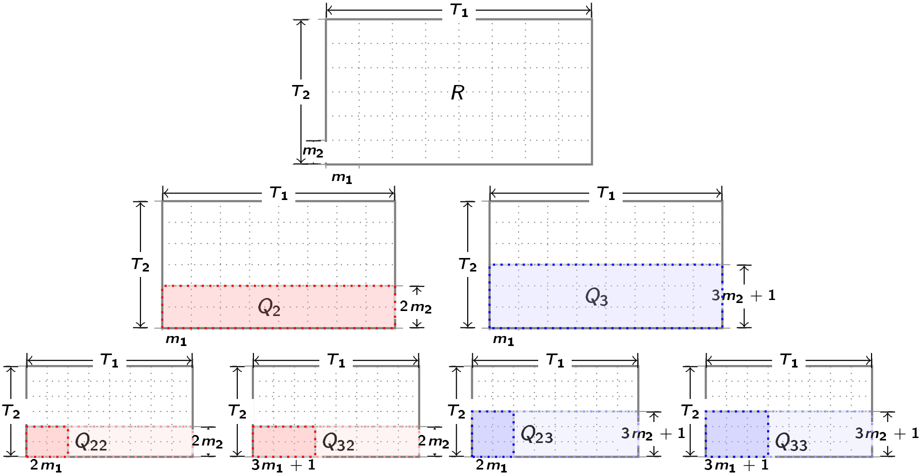
- Approximation $(1 - Q_{23} \leq \alpha_2)$

$$Q_3 \approx \frac{2Q_{23} - Q_{33}}{\left[1 + Q_{23} - Q_{33} + 2(Q_{23} - Q_{33})^2\right]^{L_1}}$$

- Error

$$L_1 E(\alpha_2, L_1)(1 - Q_{23})^2$$

Illustration of the Approximation Process



Outline

Theoretical Approximation Error

Define for $s \in \{2,3\}$

$$H(x,y,m) = \frac{2x-y}{[1+x-y+2(x-y)^2]^m}, \ \alpha_1 = 1 - Q_3, \ \alpha_2 = 1 - Q_{23},$$
$$E_1 = E(\alpha_2, L_1), \ E_2 = E(\alpha_1, L_2), \ R_s = H\left(Q_{2s}, Q_{3s}, L_1\right),$$

The approximation error

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

where B_2 is given by

$$B_2 = 1 - R_2 + L_1 F_1 (1 - Q_{22})^2$$

Outline

Simulation Error for Approximation Formula

If *ITER* is the number of simulations, we can say, at 95% confidence level,

$$\left| Q_{rt} - \hat{Q}_{rt} \right| \leq 1.96 \sqrt{\frac{\hat{Q}_{rt}(1-\hat{Q}_{rt})}{ITER}} = \beta_{rt}, \quad r,t \in \{2,3\}$$

where $\hat{Q}_{rt}$ is the simulated value.
Define for $r \in \{2,3\}$,

$$\hat{Q}_r = H\left(\hat{Q}_{2r}, \hat{Q}_{3r}, L_1\right)$$

The simulation error corresponding to the approximation formula

$$E_{sf} = L_1 L_2 \left(\beta_{22} + \beta_{23} + \beta_{32} + \beta_{33}\right)$$

Simulation Error for Approximation Error

Introducing

$$C_{2r} = 1 - \hat{Q}_{2r} + \beta_{2r}, \quad r \in \{2,3\},$$
$$C_2 = 1 - \hat{Q}_2 + L_1(\beta_{22} + \beta_{32}) + L_1 F_1 C_{22}^2,$$

The simulation error corresponding to the approximation

$$E_{sapp} = L_2 F_2 C_2^2 + L_1 L_2 F_1 \left[C_{22}^2 + C_{23}^2\right]$$

The total error

$$E_{total} = E_{app} + E_{sf} + E_{sapp}$$

Outline

Example Model

Consider for each $1 \leq i \leq T_1$ and $1 \leq j \leq T_2$:

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

- X_{ij}'s are dependent Bernoulli r.v.'s with parameter

$$p' = p\left[1 - (1-p)^8\right]$$

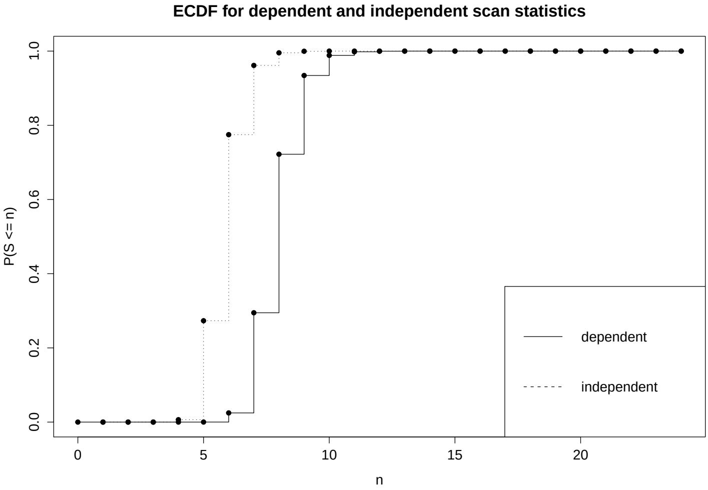
- $X_{ij} = 1$ each time when $Y_{ij} = 1$ and there is at least one success in its neighborhood (horizon one)

Numerical Results

Table 1 : $\mathbb{P}(S_{m_1,m_2}(T_1, T_2) \leq n)$: $m_1 = 4, m_2 = 6, T_1 = 53, T_2 = 75, ITER = 10^9$

n	*Sim Dep*	*Approx Dep*	E_{app}	E_{sim}	E_{total}	*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

Graphical Illustration



- Glaz, J., Naus, J., Wallenstein, S.: Scan statistic. *Springer* (2001).
- Glaz, J., Pozdnyakov, V., Wallenstein, S.: Scan statistic: Methods and Applications. *Birkhauser* (2009).
- Amarioarei, A.: *Approximation for the distribution of extremes of one dependent stationary sequences of random variables*, arXiv:1211.5456v1 (submitted)
- Amarioarei, A.: *Approximation for the Distribution of Three-dimensional Discrete Scan Statistic*, rXiv:1303.3775 (submitted)
- Boutsikas, M.V., Koutras, M.: Reliability approximations for Markov chain imbeddable systems. *Methodol Comput Appl Probab* **2** (2000), 393–412.
- 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.

- Chen, J. and Glaz, J. *Two-dimensional discrete scan statistics*, Statistics and Probability Letters 31, 59–68, 1996.
- Haiman, G.: First passage time for some stationary sequence. *Stoch Proc Appl* **80** (1999), 231–248.
- Haiman, G.: Estimating the distribution of scan statistics with high precision. *Extremes* **3** (2000), 349–361.
- Haiman, G., Preda, C.: A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. *Methodol Comput Appl Probab* **4** (2002), 393–407.
- Haiman, G., Preda, C.: Estimation for the distribution of two-dimensional scan statistics. *Methodol Comput Appl Probab* **8** (2006), 373–381.
- Haiman, G.: Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. *J. Stat Plan Infer* **137** (2007), 821–828.