Approximations for One and Two Dimensional Scan Statistics with Applications

Alexandru Amărioarei

Laboratoire de Mathématiques Paul Painlevé
Université de Lille 1, INRIA/Modal Team, France

Statistics for System Biology Seminar
November 18, 2014, Paris

Outline

Outline

One dimensional discrete scan statistics

Introducing the Model

Let $m_1 \leq T_1$ be a positive integers and $X_1, X_2, \ldots, X_{T_1}$ a sequence of r.v.'s. If we consider the moving sums

$$\boxed{Y_{i_1} = \sum_{j=i_1}^{i_1+m_1-1} X_j}$$

then the discrete one dimensional scan statistics is defined as

$$\boxed{S_{m_1}(T_1) = \max_{1 \leq i_1 \leq T_1 - m_1 + 1} Y_{i_1}.}$$

Example ($T_1 = 20,\ m_1 = 3$ and $X_{i_1} \sim \mathcal{B}(p),\ 1 \leq i_1 \leq 20$)

0	0	1	0	1	0	1	1	1	0	1	1	0	1	0	1	0	1	1	0

$Y_1 = 1$

Related statistics

Let $X_1, \ldots, X_{T_1}$ be a sequence of i.i.d. $0-1$ Bernoulli of parameter p

- $W_{m_1,k}$ - the waiting time until we first observe at least k successes in a window of size m_1

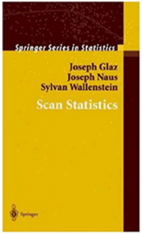
$$\mathbb{P}\left(W_{m_1,k} \leq T_1\right) = \mathbb{P}\left(S_{m_1}(T_1) \geq k\right)$$

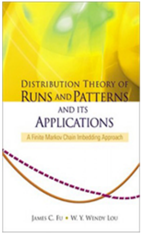
- $D_{T_1}(k)$ - the length of the smallest window that contains at least k successes

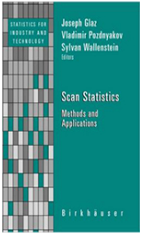
$$\mathbb{P}\left(D_{T_1}(k) \leq m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) \geq k\right)$$

- L_{T_1} - the length of the longest success run

$$\mathbb{P}\left(L_{T_1} \geq m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) \geq m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) = m_1\right)$$







Problem and approaches

Problem

Find a good estimate for the distribution of the discrete scan statistic

$$\mathbb{P}\left(S_{m_1}(T_1) \leq \tau\right).$$

Previous work:

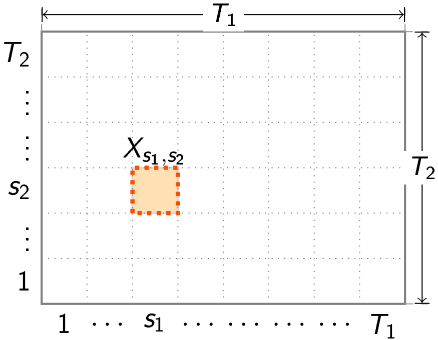
- Exact results (Bernoulli)
  - Combinatorial method: [Naus, 1974], [Naus, 1982]
  - Finite Markov chain imbedding: [Fu, 2001], [Fu and Lou, 2003], [Wu, 2013]
  - Conditional generating function: [Ebneshahrashoob and Sobel, 1990], [Gao et al., 2005]
- Approximations
  - Product-type: [Naus, 1982], [Karwe and Naus, 1997]
  - Poisson: [Chen and Glaz, 1997], [Glaz et al., 2001]
- Bounds
  - Product-type: [Glaz and Naus, 1991], [Wang et al., 2012]
  - Bonferroni: [Glaz, 1990]

Outline

Two dimensional discrete scan statistics

Introducing the Model

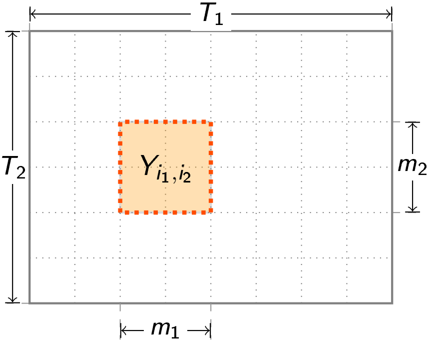
Let T_1, T_2 be positive integers



- Rectangular region $\mathcal{R}_2 = [0, T_1] \times [0, T_2]$
- $(X_{s_1,s_2})_{\substack{1\leq s_1\leq T_1\\1\leq s_2\leq T_2}}$ i.i.d. integer r.v.'s
  - Bernoulli($\mathcal{B}(1, p)$)
  - Binomial($\mathcal{B}(n, p)$)
  - Poisson($\mathcal{P}(\lambda)$)
- X_{s_1,s_2} number of observed events in the elementary subregion $r_{s_1,s_2} = [s_1 - 1, s_1] \times [s_2 - 1, s_2]$

Defining the Scan Statistic

Let m_1, m_2 be positive integers



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

$$Y_{i_1,i_2} = \sum_{s_1=i_1}^{i_1+m_1-1} \sum_{s_2=i_2}^{i_2+m_2-1} X_{s_1,s_2}$$

- The two dimensional scan statistic,

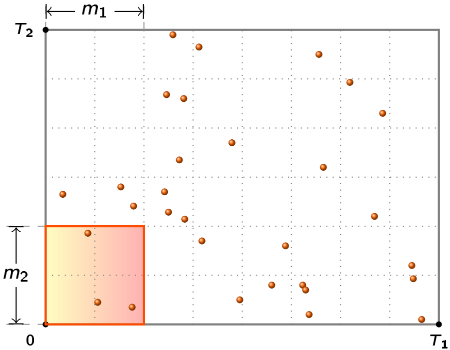
$$S_{m_1,m_2}(T_1, T_2) = \max_{\substack{1 \leq i_1 \leq T_1 - m_1 + 1 \\ 1 \leq i_2 \leq T_2 - m_2 + 1}} Y_{i_1,i_2}$$

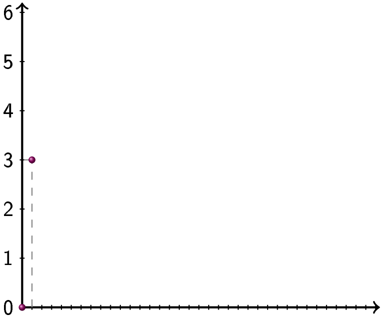
Test the null hypothesis of randomness against an alternative of clustering

H_0: The r.v.'s X_{s_1,s_2} are i.i.d. $\mathcal{B}(p)$

H_1: There exists $\mathcal{R}(i_1, i_2) = [i_1 - 1, i_1 + m_1 - 1] \times [i_2 - 1, i_2 + m_2 - 1] \subset \mathcal{R}_2$ where the r.v.'s $X_{s_1,s_2} \sim \mathcal{B}(p')$, $p' > p$ and $X_{s_1,s_2} \sim \mathcal{B}(p)$ outside $\mathcal{R}(i_1, i_2)$

Animation for 2 dimensional scan statistics





Number of points: 3

Objective

Find a good estimate for the distribution of two dimensional discrete scan statistic

$$Q_{\mathbf{m}}(\mathbf{T}) = \mathbb{P}\left(S_{\mathbf{m}}(\mathbf{T}) \leq \tau\right)$$

with $\mathbf{m} = (m_1, m_2)$ and $\mathbf{T} = (T_1, T_2)$

Previous work:

- Approximations
  - Product-type: [Boutsikas and Koutras, 2000], [Chen and Glaz, 2009]
  - Poisson: [Chen and Glaz, 1996], [Glaz et al., 2001]
- Bounds
  - Product-type (Bernoulli): [Boutsikas and Koutras, 2003]
  - Bonferroni: [Chen and Glaz, 1996], [Amărioarei, 2014]

▸ Product-Type Approximations

Outline

1. Discrete Scan Statistics (i.i.d. model)
   - One dimensional discrete scan statistics
   - Two dimensional discrete scan statistics
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples
2. Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Application: Length of the Longest increasing run
   - Application: Scan over Moving average of order q
3. Conclusions and Perspectives
4. References

Extremes of 1-dependent stationary sequences

Definitions and notations

Let $(Z_n)_{n\geq 1}$ be a sequence of random variables

m-dependence

The sequence $(Z_n)_{n\geq 1}$ is m-dependent, $m \geq 1$, if for any $h \geq 1$ the σ-fields generated by $\{Z_1, \dots, Z_h\}$ and $\{Z_{h+m+1}, \dots\}$ are independent.

Stationarity (in the strong sense)

The sequence $(Z_n)_{n\geq 1}$ is stationary if for all $k \geq 1$, for all $h \geq 0$ and for all $t_1, \dots, t_k$ the families $\{Z_{t_1}, \dots, Z_{t_k}\}$ and $\{Z_{t_1+h}, \dots, Z_{t_k+h}\}$ have the same joint distribution.

Notation

For $x < \sup\{u|\mathbb{P}(Z_1 \leq u) < 1\}$,

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

The main result

Theorem [Haiman, 1999]

For x such that $\mathbb{P}(Z_1 > x) = 1 - q_1 < 0.025$ and $n > 3$ we have

$$\left| q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n} \right| \leq n\Delta_2^H (1 - q_1)^2$$

- $\Delta_2^H = 3.3 + \frac{9}{n} + \left[15.51n(1 - q_1) + \frac{561}{n}\right](1 - q_1)$.

Theorem [Amărioarei, 2012]

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

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

- $\Delta_2 = F(q_1, n) = 1 + \frac{3}{n} + \left[K(1 - q_1) + \frac{\Gamma(1-q_1)}{n}\right](1 - q_1)$.

- Increased range of applicability
- Sharper error bounds

▸ Selected values for $K(\cdot)$ and $\Gamma(\cdot)$

The main result

Theorem [Haiman, 1999]

For x such that $\mathbb{P}(Z_1 > x) = 1 - q_1 < 0.025$ and $n > 3$ we have

$$\left| q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n} \right| \leq n\Delta_2^H (1 - q_1)^2$$

- $\Delta_2^H = 3.3 + \frac{9}{n} + \left[15.51n(1 - q_1) + \frac{561}{n}\right](1 - q_1)$.

Theorem [Amărioarei, 2012]

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

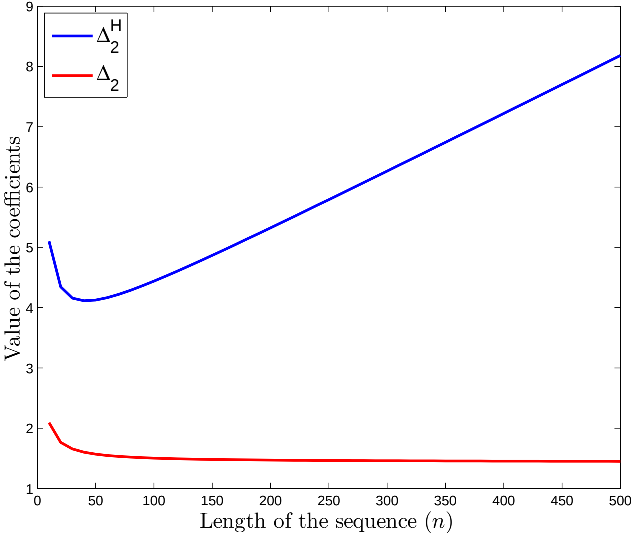
$$\left| q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n} \right| \leq n\Delta_2 (1 - q_1)^2$$

- $\Delta_2 = F(q_1, n) = 1 + \frac{3}{n} + \left[K(1 - q_1) + \frac{\Gamma(1-q_1)}{n}\right](1 - q_1)$.

- Increased range of applicability
- Sharper error bounds

Selected values for $K(\cdot)$ and $\Gamma(\cdot)$

Difference between the results: $1 - q_1 = 0.025$



Outline

1. Discrete Scan Statistics (i.i.d. model)
   - One dimensional discrete scan statistics
   - Two dimensional discrete scan statistics
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples
2. Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Application: Length of the Longest increasing run
   - Application: Scan over Moving average of order q
3. Conclusions and Perspectives
4. References

Scan statistics and 1-dependent sequences

The key idea

Main Observation

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

- The idea:
  - one dimensional scan statistic: [Haiman, 2000], [Haiman, 2007]
  - two dimensional scan statistic: [Haiman and Preda, 2002], [Haiman and Preda, 2006]
  - three dimensional scan statistic: [Amărioarei and Preda, 2013a]
  - multidimensional scan statistic: [Amărioarei, 2014]

$S_{\mathbf{m}}(\mathbf{T})$ viewed as maximum of 1-dependent r.v.'s

Let $L_j = \frac{T_j}{m_j - 1}$, $j \in \{1, 2\}$, be positive integers

- Define for each $k_1 \in \{1, 2, \ldots, L_1 - 1\}$ the random variables

$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1-1)+1 \leq i_1 \leq k_1(m_1-1) \\ 1 \leq i_2 \leq (L_2-1)(m_2-1)}} Y_{i_1, i_2}$$

- $(Z_{k_1})_{k_1}$ is 1-dependent and stationary
- Observe

$$S_{\mathbf{m}}(\mathbf{T}) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}$$

Example (One dimensional case)

$$X_1, X_2, \ldots, X_{m_1-1}, X_{m_1}, \ldots, X_{2(m_1-1)}, X_{2m_1-1}, \ldots, X_{3(m_1-1)}, X_{3m_1-2}, \ldots, X_{4(m_1-1)}$$

$S_{\mathbf{m}}(\mathbf{T})$ viewed as maximum of 1-dependent r.v.'s

Let $L_j = \frac{T_j}{m_j - 1}$, $j \in \{1, 2\}$, be positive integers

- Define for each $k_1 \in \{1, 2, \dots, L_1 - 1\}$ the random variables

$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1-1)+1 \leq i_1 \leq k_1(m_1-1) \\ 1 \leq i_2 \leq (L_2-1)(m_2-1)}} Y_{i_1, i_2}$$

- $(Z_{k_1})_{k_1}$ is 1-dependent and stationary
- Observe

$$S_{\mathbf{m}}(\mathbf{T}) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}$$

Example (One dimensional case)

$$\underbrace{X_1, X_2, \dots, X_{m_1-1}, X_{m_1}, \dots, X_{2(m_1-1)}}_{Z_1}, X_{2m_1-1}, \dots, X_{3(m_1-1)}, X_{3m_1-2}, \dots, X_{4(m_1-1)}$$

$S_{\mathbf{m}}(\mathbf{T})$ viewed as maximum of 1-dependent r.v.'s

Let $L_j = \frac{T_j}{m_j - 1}$, $j \in \{1, 2\}$, be positive integers

- Define for each $k_1 \in \{1, 2, \ldots, L_1 - 1\}$ the random variables

$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1-1)+1 \leq i_1 \leq k_1(m_1-1) \\ 1 \leq i_2 \leq (L_2-1)(m_2-1)}} Y_{i_1, i_2}$$

- $(Z_{k_1})_{k_1}$ is 1-dependent and stationary
- Observe

$$S_{\mathbf{m}}(\mathbf{T}) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}$$

Example (One dimensional case)

$$\underbrace{X_1, X_2, \ldots, X_{m_1-1}, \overbrace{X_{m_1}, \ldots, X_{2(m_1-1)}}}_{Z_1}{}^{\!\!\!\!Z_2\;\;} \overbrace{X_{2m_1-1}, \ldots, X_{3(m_1-1)}}^{Z_2}, X_{3m_1-2}, \ldots, X_{4(m_1-1)}$$

$S_{\mathbf{m}}(\mathbf{T})$ viewed as maximum of 1-dependent r.v.'s

Let $L_j = \frac{T_j}{m_j - 1}$, $j \in \{1, 2\}$, be positive integers

- Define for each $k_1 \in \{1, 2, \dots, L_1 - 1\}$ the random variables

$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1-1)+1 \leq i_1 \leq k_1(m_1-1) \\ 1 \leq i_2 \leq (L_2-1)(m_2-1)}} Y_{i_1, i_2}$$

- $(Z_{k_1})_{k_1}$ is 1-dependent and stationary
- Observe

$$S_{\mathbf{m}}(\mathbf{T}) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}$$

Example (One dimensional case)

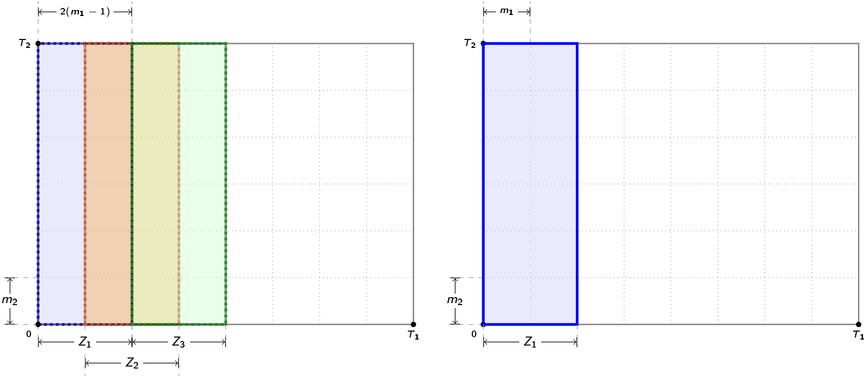
$$\underbrace{X_1, X_2, \dots, X_{m_1-1}, \overbrace{X_{m_1}, \dots, X_{2(m_1-1)}}_{}}_{Z_1}\overbrace{, X_{2m_1-1}, \dots, X_{3(m_1-1)}}^{Z_2}, \dots$$

$$X_1, X_2, \dots, X_{m_1-1}, \overbrace{X_{m_1}, \dots, X_{2(m_1-1)}, X_{2m_1-1}, \dots, X_{3(m_1-1)}}^{Z_2}, X_{3m_1-2}, \dots, X_{4(m_1-1)}$$

$$\underbrace{X_1, X_2, \dots, X_{m_1-1}, X_{m_1}, \dots, X_{2(m_1-1)}}_{Z_1}, \underbrace{X_{2m_1-1}, \dots, X_{3(m_1-1)}, X_{3m_1-2}, \dots, X_{4(m_1-1)}}_{Z_3}$$

$S_m(T)$ viewed as maximum of 1-dependent r.v.'s

Example (Two dimensional case)



Approximation process: First step

Define for $t_1 \in \{2, 3\}$,

$$Q_{t_1} = Q_{t_1}(\tau) = \mathbb{P}\left(\bigcap_{k_1=1}^{t_1-1}\{Z_{k_1} \leq \tau\}\right) = \mathbb{P}\left(\max_{\substack{1\leq i_1\leq (t_1-1)(m_1-1)\\ 1\leq i_2\leq (L_2-1)(m_2-1)}} Y_{i_1,i_2} \leq \tau\right)$$

If $1 - Q_2 \leq 0.1$ then

$$\left|Q_m(\mathbf{T}) - \frac{2Q_2 - Q_3}{[1+Q_2-Q_3+2(Q_2-Q_3)^2]^{L_1-1}}\right| \leq (L_1-1)F(Q_2, L_1-1)(1-Q_2)^2$$

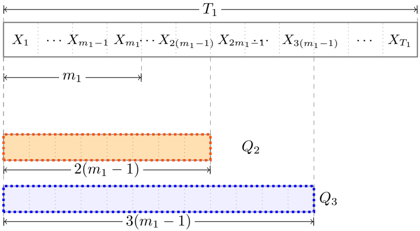
Example (One dimensional case)

- The approximation

$$\mathbb{P}(S_{m_1}(T_1) \leq \tau) \approx \frac{2Q_2 - Q_3}{[1+Q_2-Q_3+2(Q_2-Q_3)^2]^{L_1-1}}$$

- Approximation error, about

$$(L_1-1)F(Q_2, L_1-1)\,(1-Q_2)^2$$



Approximation process: Second step

The approximation of $S_{\mathbf{m}}(\mathbf{T})$ is an iterative process. The second step becomes:

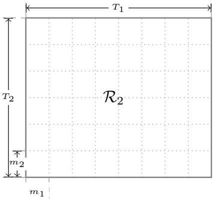
- Define for $t_1 \in \{2,3\}$ and $k_2 \in \{1,2,\ldots,L_2-1\}$

$$\boxed{Z_{k_2}^{(t_1)} = \max_{\substack{1 \leq i_1 \leq (t_1-1)(m_1-1) \\ (k_2-1)(m_2-1)+1 \leq i_2 \leq k_2(m_2-1)}} Y_{i_1,i_2}}$$

- $\left\{Z_1^{(t_1)},\ldots,Z_{L_2-1}^{(t_1)}\right\}$ forms a 1-dependent stationary sequence
- If we take $H(x,y,m) = \frac{2x-y}{[1+x-y+2(x-y)^2]^{m-1}}$, then we have the approximation

$$\boxed{|Q_{t_1} - H\left(Q_{t_1,2}, Q_{t_1,3}, L_2\right)| \leq (L_2-1)F(Q_{t_1,2}, L_2-1)(1-Q_{t_1,2})^2}$$

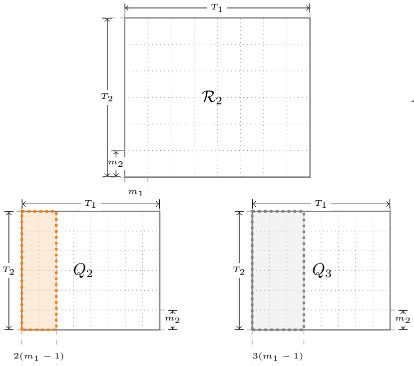
Illustration for the two dimensional case



Find

Approximation

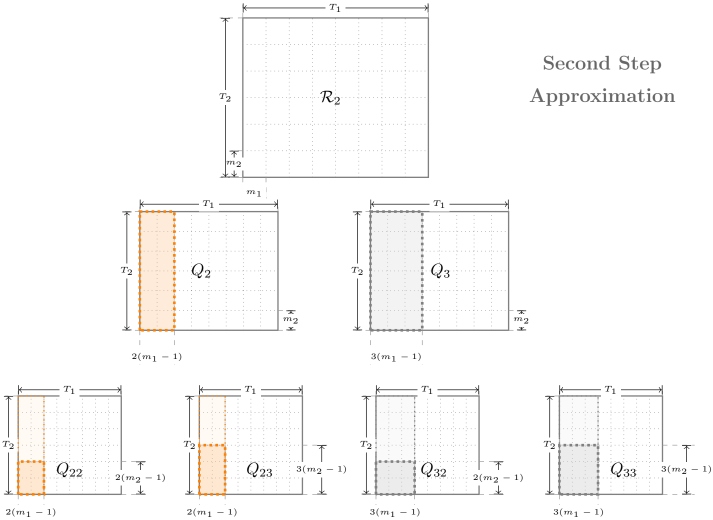
Illustration for the two dimensional case



First Step

Approximation

Illustration for the two dimensional case



Second Step

Approximation

Error bounds

Let $\gamma_{t_1,t_2} = Q_{t_1,t_2}$, with $t_j \in \{2,3\}$, $j \in \{1,2\}$, and define

$$\gamma_{t_1} = H\left(\gamma_{t_1,2}, \gamma_{t_1,3}, L_2\right)$$

Denote with $\hat{Q}_{t_1,t_2}$ the estimated value of Q_{t_1,t_2} and define

$$\hat{Q}_{t_1} = H\left(\hat{Q}_{t_1,2}, \hat{Q}_{t_1,3}, L_2\right)$$

Objective

$$\mathbf{Q_m(T) \approx H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)}$$

We observe that

$$\left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right| \leq \left|Q_{\mathbf{m}}(\mathbf{T}) - H\left(\gamma_2, \gamma_3, L_1\right)\right| + \left|H\left(\gamma_2, \gamma_3, L_1\right) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right|$$

The quantities $\hat{Q}_{t_1,t_2}$ will be estimated by Monte Carlo simulations. ▸ Error bounds

Université Lille1 Sciences et Technologies

Error bounds

Let $\gamma_{t_1,t_2} = Q_{t_1,t_2}$, with $t_j \in \{2,3\}$, $j \in \{1,2\}$, and define

$$\gamma_{t_1} = H\left(\gamma_{t_1,2}, \gamma_{t_1,3}, L_2\right)$$

Denote with $\hat{Q}_{t_1,t_2}$ the estimated value of Q_{t_1,t_2} and define

$$\hat{Q}_{t_1} = H\left(\hat{Q}_{t_1,2}, \hat{Q}_{t_1,3}, L_2\right)$$

Objective

$$\mathbf{Q_m(T) \approx H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)}$$

We observe that

$$\left|Q_m(\mathbf{T}) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right| \leq \underbrace{\left|Q_m(\mathbf{T}) - H\left(\gamma_2, \gamma_3, L_1\right)\right|}_{E_{app}(2)} + \underbrace{\left|H\left(\gamma_2, \gamma_3, L_1\right) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right|}_{E_{sf}(2)}$$

The quantities $\hat{Q}_{t_1,t_2}$ will be estimated by Monte Carlo simulations. Error bounds

Université Lille1 Sciences et Technologies

Error bounds

Let $\gamma_{t_1,t_2} = Q_{t_1,t_2}$, with $t_j \in \{2,3\}$, $j \in \{1,2\}$, and define

$$\gamma_{t_1} = H\left(\gamma_{t_1,2}, \gamma_{t_1,3}, L_2\right)$$

Denote with $\hat{Q}_{t_1,t_2}$ the estimated value of Q_{t_1,t_2} and define

$$\hat{Q}_{t_1} = H\left(\hat{Q}_{t_1,2}, \hat{Q}_{t_1,3}, L_2\right)$$

Objective

$$\mathbf{Q_m(T) \approx H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)}$$

We observe that

$$\left|Q_m(\mathbf{T}) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right| \leq \underbrace{\left|Q_m(\mathbf{T}) - H\left(\gamma_2, \gamma_3, L_1\right)\right|}_{E_{app}(2) \leq E_{sapp}(2)} + \underbrace{\left|H\left(\gamma_2, \gamma_3, L_1\right) - H\left(\hat{Q}_2, \hat{Q}_3, L_1\right)\right|}_{E_{sf}(2)}$$

The quantities $\hat{Q}_{t_1,t_2}$ will be estimated by Monte Carlo simulations. ▸ Error bounds

Outline

Simulation methods and computational aspects

Naive Hit-or-Miss MC

Objective

Find an estimate for $\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right)$.

Algorithm 1 Classical Monte Carlo algorithm for scan statistics

Begin

Repeat for each k from 1 to $ITER$ (iterations number)

1: Generate $\mathbf{X}^{(\mathbf{k})} = \left\{X_{s_1,s_2}^{(k)}, 1 \leq s_j \leq T_j, 1 \leq j \leq 2\right\}$ under H_0

2: Compute the two dimensional scan statistics $S_{\mathbf{m}}^{(k)}(\mathbf{T})$ over $\mathbf{X}^{(\mathbf{k})}$

End Repeat

Return

$$\widehat{p_{MC}} = \frac{1}{ITER}\sum_{i=1}^{ITER} \mathbf{1}_{\left\{S_{\mathbf{m}}^{(i)}(\mathbf{T}) \geq \tau\right\}}, \quad \widehat{s.e._{MC}} = \sqrt{\frac{\widehat{p_{MC}}(1-\widehat{p_{MC}})}{ITER}}$$

the unbiased direct Monte Carlo estimate and its consistent standard error estimate.

End

- computationally intensive since just a fraction of the generated observations will cause a rejection
- needs a large number of replications in order to reduce the standard error estimate to an acceptable level

Importance sampling for scan statistics

Idea behind importance sampling

Find a good change of measure that leads to an efficient sampling process.

The method was previously used for solving the problem of:

- union count: [Frigessi and Vercellis, 1984], [Fishman, 1996]
- exceeding probabilities: [Naiman and Wynn, 1997]
- scan statistics: [Naiman and Priebe, 2001], [Malley et al., 2002]

We are interested in evaluating the probability

$$\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) = \mathbb{P}\left(\bigcup_{i_1=1}^{T_1-m_1+1}\bigcup_{i_2=1}^{T_2-m_2+1} E_{i_1,i_2}\right) = \int G(\mathbf{x})f(\mathbf{x})\,d\mathbf{x}$$

where $E_{i_1,i_2} = \{Y_{i_1,i_2} \geq \tau\}$, $G(\mathbf{x}) = \mathbf{1}_E(\mathbf{x})$, $E = \bigcup_{i_1=1}^{T_1-m_1+1}\bigcup_{i_2=1}^{T_2-m_2+1} E_{i_1,i_2}$ and f is the joint density of Y_{i_1,i_2} under H_0.

Importance sampling for scan statistics

We introduce the change of measure

$$g(\mathbf{x}) = \sum_{j_1=1}^{T_1-m_1+1} \sum_{j_2=1}^{T_2-m_2+1} \left\{ \frac{\mathbb{P}\left(E_{j_1,j_2}\right)}{B(2)} \right\} \left\{ \frac{\mathbf{1}_{E_{j_1,j_2}} f(\mathbf{x})}{\mathbb{P}\left(E_{j_1,j_2}\right)} \right\}$$

and we observe that $\mathbb{P}_{H_0}\left(S_{\mathbf{m}}(\mathbf{T}) \geq \tau\right) = B(2)\rho(2)$

- the Bonferroni upper bound $B(2)$

$$B(d) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} \mathbb{P}\left(E_{i_1,i_2}\right)$$

- the correction factor $\rho(2)$ between 0 and 1

$$\rho(2) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} p_{i_1,i_2} \int \frac{1}{C(\mathbf{Y})} d\mathbb{P}_{H_0}(\cdot|E_{i_1,i_2})$$

where

$$p_{i_1,i_2} = \frac{1}{(T_1-m_1+1)(T_2-m_2+1)}, \quad C(\mathbf{Y}) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} \mathbf{1}_{E_{i_1,i_2}}$$

Importance sampling for scan statistics

Algorithm 2 Importance Sampling Algorithm for Scan Statistics

Begin
Repeat for each k from 1 to $ITER$ (iterations number)

1: Generate uniformly the couple $(i_1^{(k)}, i_2^{(k)})$ from the set $\{1,\dots,T_1-m_1+1\}\times\{1,\dots,T_2-m_2+1\}$.

2: Given the couple $(i_1^{(k)}, i_2^{(k)})$, generate a sample of the random field $\tilde{\mathbf{X}}^{(k)} = \left\{\tilde{X}_{s_1,s_2}^{(k)}\right\}$, with $s_j \in \{1,\dots,T_j\}$ and $j\in\{1,2\}$, from the conditional distribution of $\mathbf{X}$ given $\left\{Y_{i_1^{(k)},i_2^{(k)}} \geq \tau\right\}$.

3: Take $c_k = C(\tilde{\mathbf{X}}^{(k)})$ the number of all couples (i_1,i_2) for which $\tilde{Y}_{i_1,i_2}\geq\tau$ and put $\widehat{\rho}_k(2) = \frac{1}{c_k}$.

End Repeat
Return

$$\widehat{\rho}(2) = \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(2), \quad Var\left[\widehat{\rho}(2)\right] \approx \frac{1}{ITER-1}\sum_{k=1}^{ITER}\left(\widehat{\rho}_k(2) - \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(2)\right)^2$$

End

Implementation problems

Algorithm 2 presents two main difficulties:

A) being able to sample from the conditional distribution of $\mathbf{X}$ given $\left\{Y_{i_1^{(k)},i_2^{(k)}} \geq \tau\right\}$ in **Step** 2

B) the number of locality statistics that exceed the predetermined threshold is supposed to be found in a *reasonable* time

Partial solutions were found for:

A) binomial, Poisson and Gaussian model

B) *cumulative counts* or *fast spatial scan* techniques (see [Neil, 2006], [Neil, 2012])

Scan 1d for normal data

Outline

Numerical examples

One dimensional case: $X_{i_1} \sim \mathcal{B}(n, p)$

Table 1 : $n = 1, p = 0.005, m_1 = 10, T_1 = 1000, It_{App} = 10^4$

τ	Exact	Glaz et al. Product-type	Our Approximation	Approximation Error	Lower Bound	Upper Bound
1	0.810209	0.810216	0.810404	0.001111	0.809903	0.810439
2	0.995764	0.995764	0.995764	3×10^{-7}	0.995764	0.995764
3	0.999950	0.999950	0.999950	4×10^{-11}	0.999950	0.999950

Table 2 : $n = 5, p = 0.05, m_1 = 25, T_1 = 500, It_{App} = 10^4, It_{Sim} = 10^3$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
13	0.712750	0.705787	0.714699	0.039308	0.697431	0.706948
14	0.867498	0.862184	0.865029	0.012502	0.859543	0.862407
15	0.946912	0.943329	0.946177	0.004169	0.942552	0.943362
16	0.980230	0.978959	0.979822	0.001354	0.978733	0.978963
17	0.993486	0.992821	0.993134	0.000433	0.992756	0.992822
18	0.997802	0.997726	0.997849	0.000127	0.997708	0.997726
19	0.999362	0.999327	0.999358	3×10^{-5}	0.999322	0.999327
20	0.999819	0.999813	0.999825	9×10^{-6}	0.999812	0.999813
21	0.999954	0.999951	0.999953	2×10^{-6}	0.999951	0.999951

Other numerical results

One dimensional case: $X_{i_1} \sim \mathcal{B}(n, p)$

Table 1 : $n = 1, p = 0.005, m_1 = 10, T_1 = 1000, It_{App} = 10^4$

τ	Exact	Glaz et al. Product-type	Our Approximation	Approximation Error	Lower Bound	Upper Bound
1	0.810209	0.810216	0.810404	0.001111	0.809903	0.810439
2	0.995764	0.995764	0.995764	3×10^{-7}	0.995764	0.995764
3	0.999950	0.999950	0.999950	4×10^{-11}	0.999950	0.999950

Table 2 : $n = 5, p = 0.05, m_1 = 25, T_1 = 500, It_{App} = 10^4, It_{Sim} = 10^3$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
13	0.712750	0.705787	0.714699	0.039308	0.697431	0.706948
14	0.867498	0.862184	0.865029	0.012502	0.859543	0.862407
15	0.946912	0.943329	0.946177	0.004169	0.942552	0.943362
16	0.980230	0.978959	0.979822	0.001354	0.978733	0.978963
17	0.993486	0.992821	0.993134	0.000433	0.992756	0.992822
18	0.997802	0.997726	0.997849	0.000127	0.997708	0.997726
19	0.999362	0.999327	0.999358	3×10^{-5}	0.999322	0.999327
20	0.999819	0.999813	0.999825	9×10^{-6}	0.999812	0.999813
21	0.999954	0.999951	0.999953	2×10^{-6}	0.999951	0.999951

▸ Other numerical results

One dimensional case: $X_{i_1} \sim \mathcal{B}(n, p)$

Table 1 : $n = 1, p = 0.005, m_1 = 10, T_1 = 1000, It_{App} = 10^4$

τ	Exact	Glaz et al. Product-type	Our Approximation	Approximation Error	Lower Bound	Upper Bound
1	0.810209	0.810216	0.810404	0.001111	0.809903	0.810439
2	0.995764	0.995764	0.995764	3×10^{-7}	0.995764	0.995764
3	0.999950	0.999950	0.999950	4×10^{-11}	0.999950	0.999950

Table 2 : $n = 5, p = 0.05, m_1 = 25, T_1 = 500, It_{App} = 10^4, It_{Sim} = 10^3$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
13	0.712750	0.705787	0.714699	0.039308	0.697431	0.706948
14	0.867498	0.862184	0.865029	0.012502	0.859543	0.862407
15	0.946912	0.943329	0.946177	0.004169	0.942552	0.943362
16	0.980230	0.978959	0.979822	0.001354	0.978733	0.978963
17	0.993486	0.992821	0.993134	0.000433	0.992756	0.992822
18	0.997802	0.997726	0.997849	0.000127	0.997708	0.997726
19	0.999362	0.999327	0.999358	3×10^{-5}	0.999322	0.999327
20	0.999819	0.999813	0.999825	9×10^{-6}	0.999812	0.999813
21	0.999954	0.999951	0.999953	2×10^{-6}	0.999951	0.999951

Other numerical results

Two dimensional case: $X_{i_1,i_2} \sim \mathcal{B}(n,p)$

Table 3 : $n = 1, p = 0.005, m_1 = m_2 = 6, T_1 = T_2 = 30, It_{App} = 10^3, It_{Sim} = 10^3$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
2	0.915903	0.914013	0.920211	0.041483	0.901935	0.945623
3	0.994292	0.994395	0.994578	0.000803	0.993785	0.996638
4	0.999747	0.999757	0.999760	2×10^{-5}	0.999737	0.999858
5	0.999992	0.999992	0.999992	7×10^{-7}	0.999992	0.999995

Table 4 : $n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, It_{App} = 10^4$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
4	0.894654	0.873256	0.893724	0.037136	0.803422	0.944318
5	0.988003	0.986249	0.988144	0.002125	0.981418	0.993451
6	0.998963	0.998847	0.998963	0.000152	0.998543	0.999401
7	0.999926	0.999919	0.999925	9×10^{-6}	0.999903	0.999955
8	0.999995	0.999995	0.999995	5×10^{-7}	0.999994	0.999997

Two dimensional case: $X_{i_1,i_2} \sim \mathcal{B}(n,p)$

Table 3 : $n = 1, p = 0.005, m_1 = m_2 = 6, T_1 = T_2 = 30, It_{App} = 10^3, It_{Sim} = 10^3$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
2	0.915903	0.914013	0.920211	0.041483	0.901935	0.945623
3	0.994292	0.994395	0.994578	0.000803	0.993785	0.996638
4	0.999747	0.999757	0.999760	2×10^{-5}	0.999737	0.999858
5	0.999992	0.999992	0.999992	7×10^{-7}	0.999992	0.999995

Table 4 : $n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, It_{App} = 10^4$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
4	0.894654	0.873256	0.893724	0.037136	0.803422	0.944318
5	0.988003	0.986249	0.988144	0.002125	0.981418	0.993451
6	0.998963	0.998847	0.998963	0.000152	0.998543	0.999401
7	0.999926	0.999919	0.999925	9×10^{-6}	0.999903	0.999955
8	0.999995	0.999995	0.999995	5×10^{-7}	0.999994	0.999997

Two dimensional case: $X_{i_1,i_2} \sim \mathcal{B}(n,p)$

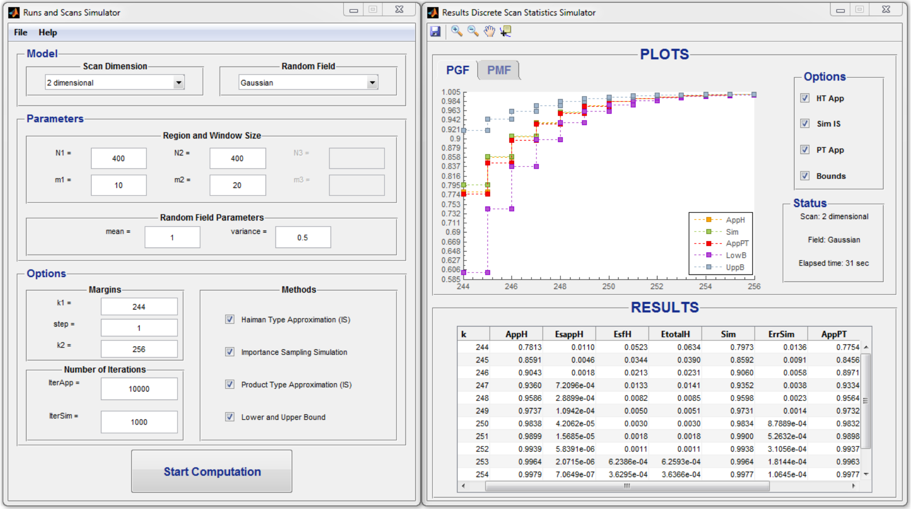
Table 3 : $n=1, p=0.005, m_1=m_2=6, T_1=T_2=30, It_{App}=10^3, It_{Sim}=10^3$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
2	0.915903	0.914013	0.920211	0.041483	0.901935	0.945623
3	0.994292	0.994395	0.994578	0.000803	0.993785	0.996638
4	0.999747	0.999757	0.999760	2×10^{-5}	0.999737	0.999858
5	0.999992	0.999992	0.999992	7×10^{-7}	0.999992	0.999995

Table 4 : $n=5, p=0.002, m_1=5, m_2=10, T_1=50, T_2=80, It_{App}=10^4$

τ	$\hat{\mathbb{P}}(S \leq \tau)$	Glaz et al. Product-type	Our Approximation	Total Error	Lower Bound	Upper Bound
4	0.894654	0.873256	0.893724	0.037136	0.803422	0.944318
5	0.988003	0.986249	0.988144	0.002125	0.981418	0.993451
6	0.998963	0.998847	0.998963	0.000152	0.998543	0.999401
7	0.999926	0.999919	0.999925	9×10^{-6}	0.999903	0.999955
8	0.999995	0.999995	0.999995	5×10^{-7}	0.999994	0.999997

Matlab GUI application



k	AppH	EsappH	EsfH	EtotalH	Sim	ErrSim	AppPT
244	0.7813	0.0110	0.0523	0.0634	0.7973	0.0136	0.7754
245	0.8591	0.0046	0.0344	0.0390	0.8592	0.0091	0.8456
246	0.9043	0.0018	0.0213	0.0231	0.9060	0.0058	0.8971
247	0.9360	7.2096e-04	0.0133	0.0141	0.9352	0.0038	0.9334
248	0.9586	2.8899e-04	0.0082	0.0085	0.9598	0.0023	0.9564
249	0.9737	1.0942e-04	0.0050	0.0051	0.9731	0.0014	0.9732
250	0.9838	4.2062e-05	0.0030	0.0030	0.9834	8.7889e-04	0.9832
251	0.9899	1.5685e-05	0.0018	0.0018	0.9900	5.2632e-04	0.9898
252	0.9939	5.8391e-06	0.0011	0.0011	0.9938	3.1056e-04	0.9937
253	0.9964	2.0715e-06	6.2386e-04	6.2593e-04	0.9964	1.8144e-04	0.9963
254	0.9979	7.0649e-07	3.6295e-04	3.6366e-04	0.9977	1.0645e-04	0.9977

Outline

Model and discussion

Definition of a block-factor

k block-factor

The sequence $(Z_n)_{n\geq 1}$ of random variables with state space S_W is said to be k block-factor of the sequence $(Y_n)_{n\geq 1}$ with state space S_Y if there is a measurable function $f : S_Y^k \to S_W$ such that

$$Z_n = f\left(Y_n, Y_{n+1}, \dots, Y_{n+k-1}\right), \forall n \geq 1.$$

Example (2 block-factors)

- $Z_n = Y_n + Y_{n+1},\ n \geq 1$ for $f(x, y) = x + y$
- $Z_n = Y_n Y_{n+1},\ n \geq 1$ for $f(x, y) = xy$

Observation

If a sequence $(Z_n)_{n\geq 1}$ of random variables is a k block-factor, then the sequence is $(k-1)$-dependent.

Introducing the model

For each $1 \leq j \leq 2$, let $\tilde{T}_j$, $x_1^{(j)}$, $x_2^{(j)}$, $c_j = x_1^{(j)} + x_2^{(j)} + 1$, $T_j = \tilde{T}_j - c_j + 1$ and $2 \leq m_j \leq T_j$ be nonnegative integers.

- The rectangular region, $\tilde{\mathcal{R}}_2 = [0, \tilde{T}_1] \times [0, \tilde{T}_2]$
- $\tilde{X}_{s_1,s_2}$, $1 \leq s_j \leq \tilde{T}_j$, $j \in \{1, 2\}$ be i.i.d. r.v.'s

To each couple (s_1, s_2), with $s_j \in \left\{x_1^{(j)} + 1, \dots, \tilde{T}_j - x_2^{(j)}\right\}$, $j \in \{1, 2\}$, associate a 2-way tensor (matrix) $\mathcal{X}_{s_1,s_2} \in \mathbb{R}^{c_1 \times c_2}$

$$\mathcal{X}_{s_1,s_2}(j_1, j_2) = \tilde{X}_{s_1 - x_1^{(1)} - 1 + j_1,\, s_2 - x_1^{(2)} - 1 + j_2}$$

where $(j_1, j_2) \in \{1, \dots, c_1\} \times \{1, \dots, c_2\}$.

Let $\Pi : \mathbb{R}^{c_1 \times c_2} \to \mathbb{R}$ be a measurable real valued function and define, for all $1 \leq s_j \leq T_j$, $1 \leq j \leq 2$, the *block-factor type* model

$$X_{s_1,s_2} = \Pi\left(\mathcal{X}_{s_1 + x_1^{(1)},\, s_2 + x_1^{(2)}}\right)$$

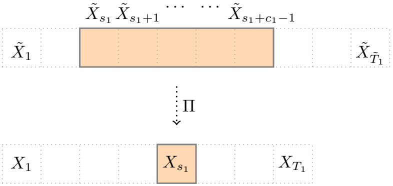
- [Amărioarei and Preda, 2013b] and [Amărioarei and Preda, 2014]

Examples for one and two dimensions

Example (One dimensional case)

$$x_{s_1} = \left[\tilde{X}_{s_1 - x_1^{(1)}}, \dots, \tilde{X}_{s_1 + x_2^{(1)}}\right]$$

$$X_{s_1} = \Pi\left(x_{s_1 + x_1^{(1)}}\right) = \Pi\left(\tilde{X}_{s_1}, \dots, \tilde{X}_{s_1 + c_1 - 1}\right)$$

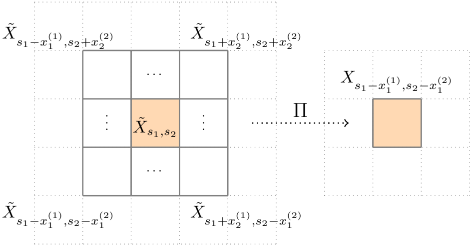


Example (Two dimensional case)

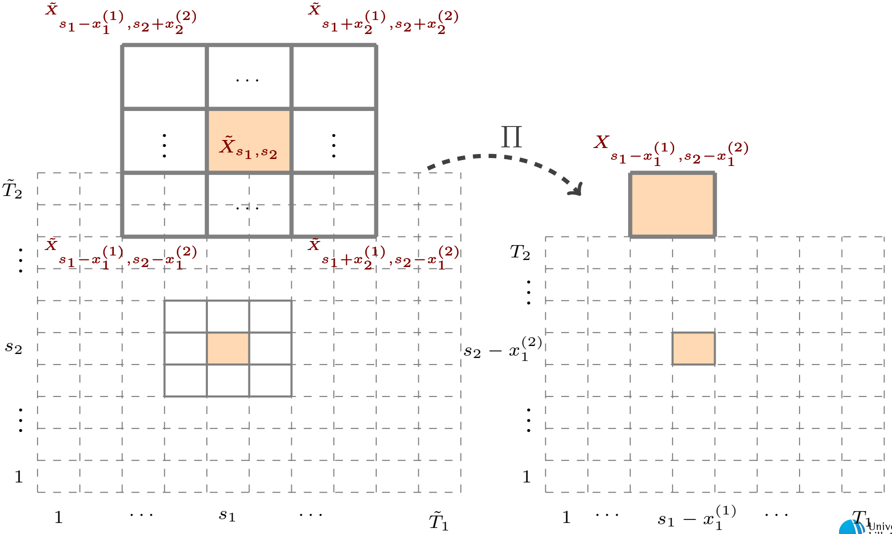
$$x_{s_1,s_2} =$$

$$\begin{pmatrix} \tilde{X}_{s_1 - x_1^{(1)}, s_2 - x_1^{(2)}} & \cdots & \tilde{X}_{s_1 + x_2^{(1)}, s_2 - x_1^{(2)}} \\ \vdots & \ddots & \vdots \\ \tilde{X}_{s_1 - x_1^{(1)}, s_2 + x_2^{(2)}} & \cdots & \tilde{X}_{s_1 + x_2^{(1)}, s_2 + x_2^{(2)}} \end{pmatrix}$$

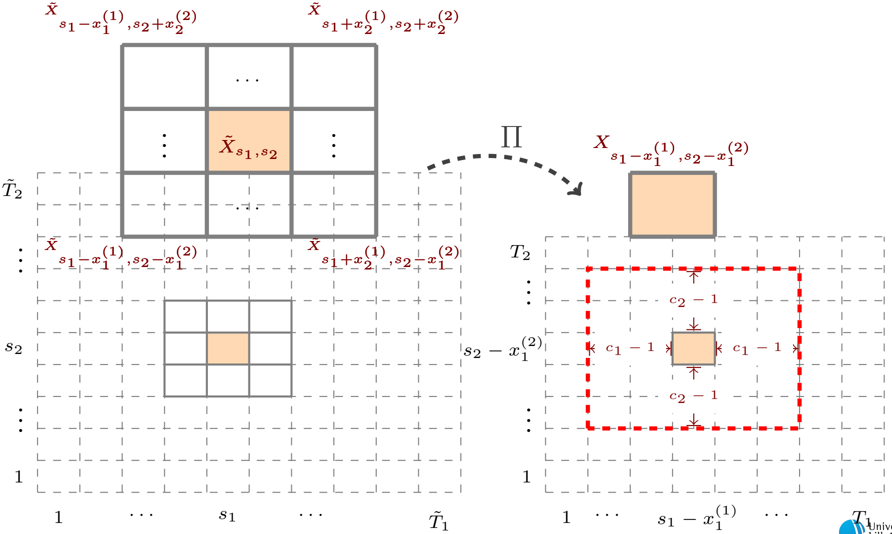
$$X_{s_1,s_2} = \Pi\left(x_{s_1 + x_1^{(1)}, s_2 + x_1^{(2)}}\right)$$



Dependency structure in two dimensions



Dependency structure in two dimensions



Approximation: idea

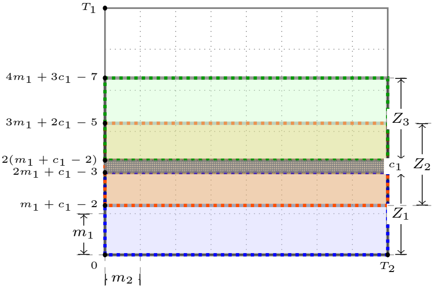
Let $L_j = \frac{\tilde{T}_j}{m_j+c_j-2}$, $j \in \{1,2\}$, be positive integers

- Define for each $k_1 \in \{1,2,\dots,L_1-1\}$ the random variables

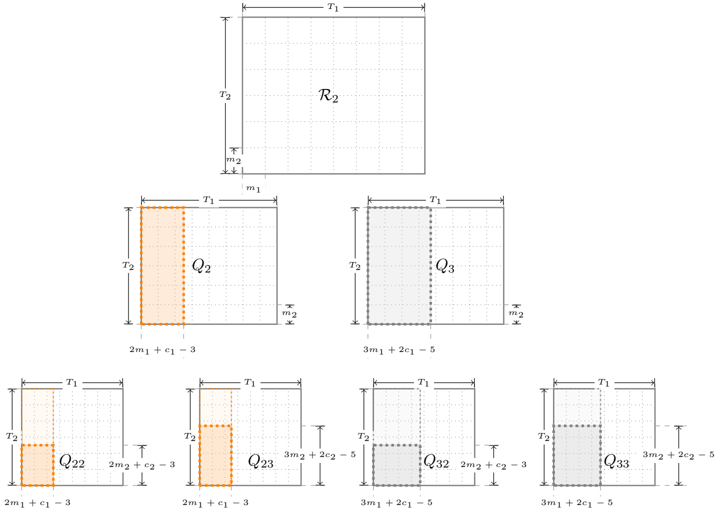
$$Z_{k_1} = \max_{\substack{(k_1-1)(m_1+c_1-2)+1 \leq i_1 \leq k_1(m_1+c_1-2)\\ 1 \leq i_2 \leq (L_2-1)(m_2+c_2-2)}} Y_{i_1,i_2}$$

- $(Z_{k_1})_{k_1}$ is 1-dependent, stationary and $S_{\mathbf{m}}(\mathbf{T}) = \max_{1\leq k_1 \leq L_1-1} Z_{k_1}$

Illustration of the 1-dependence structure in two dimensions



Approximation process in two dimensions



Error bounds

Outline

1. Discrete Scan Statistics (i.i.d. model)
   - One dimensional discrete scan statistics
   - Two dimensional discrete scan statistics
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples
2. Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Application: Length of the Longest increasing run
   - Application: Scan over Moving average of order q
3. Conclusions and Perspectives
4. References

Application

Longest increasing run

Let $\left(\tilde{X}_n\right)_{n\geq 1}$ be a sequence of i.i.d. r.v.'s with the common distribution G.

Increasing run

A subsequence $\left(\tilde{X}_k, \dots, \tilde{X}_{k+l-1}\right)$ forms an *increasing run* of length $l \geq 1$, starting at position $k \geq 1$, if

$$\tilde{X}_{k-1} > \tilde{X}_k < \tilde{X}_{k+1} < \cdots < \tilde{X}_{k+l-1} > \tilde{X}_{k+l}$$

Notations

- $M_{\tilde{T}_1}$ = the length of the longest increasing run among the first $\tilde{T}_1$ r.v.'s
- $L_{\tilde{T}_1}$ = the length of the longest run of ones among the first $\tilde{T}_1$ r.v.'s

The asymptotic distribution was studied

- G continuous distribution: [Pittel, 1981], [Révész, 1983], [Grill, 1987], [Novak, 1992], etc.
- G discrete distribution: [Csaki and Foldes, 1996], [Grabner et al., 2003], [Eryilmaz, 2006], etc.

Longest increasing run

Scan statistics approach

In the one dimensional problem, let $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ and define $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

$$\Pi(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

Example $(\tilde{X}_{s_1} \sim \mathcal{U}(0,1),\ \tilde{T}_1 = 10)$

$\tilde{X}_{s_1}$:	0.79	0.31	0.52	0.16	0.60	0.26	0.65	0.68	0.74	0.45
X_{s_1} :										

We have

$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right), \text{ for } m_1 \geq 1$$

Longest increasing run

Scan statistics approach

In the one dimensional problem, let $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ and define $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

$$\Pi(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

Example $(\tilde{X}_{s_1} \sim \mathcal{U}(0,1),\ \tilde{T}_1 = 10)$

$\tilde{X}_{s_1}$: 0.79 0.31 0.52 0.16 0.60 0.26 0.65 0.68 0.74 0.45

X_{s_1} : 0

We have

$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right), \text{ for } m_1 \geq 1$$

Longest increasing run

Scan statistics approach

In the one dimensional problem, let $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ and define $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

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

- the block-factor model becomes: $X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

Example $(\tilde{X}_{s_1} \sim \mathcal{U}(0,1),\ \tilde{T}_1 = 10)$

$$\begin{array}{lcccccccccc} \tilde{X}_{s_1}: & 0.79 & 0.31 & 0.52 & 0.16 & 0.60 & 0.26 & 0.65 & 0.68 & 0.74 & 0.45 \\ & \searrow & \swarrow\ \searrow & \swarrow & & & & & & & \\ X_{s_1}: & & 0 \quad 1 & & & & & & & & \end{array}$$

We have

$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right), \text{ for } m_1 \geq 1$$

Longest increasing run

Scan statistics approach

In the one dimensional problem, let $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ and define $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

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

- the block-factor model becomes: $X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

Example $(\tilde{X}_{s_1} \sim \mathcal{U}(0,1),\ \tilde{T}_1 = 10)$

$$\begin{array}{lcccccccccccccccccccc}
\tilde{X}_{s_1}: & 0.79 & & 0.31 & & 0.52 & & 0.16 & & 0.60 & & 0.26 & & 0.65 & & 0.68 & & 0.74 & & 0.45 \\
 & & \searrow\swarrow & & \searrow\swarrow & & \searrow\swarrow & & & & & & & & & & & & & \\
X_{s_1}: & & 0 & & 1 & & 0 & & & & & & & & & & & & &
\end{array}$$

We have

$$\mathbb{P}\left(M_{\tilde{T}_1} \le m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right), \text{ for } m_1 \ge 1$$

Longest increasing run

Scan statistics approach

In the one dimensional problem, let $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ and define $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

$$\Pi(x, y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

Example $(\tilde{X}_{s_1} \sim \mathcal{U}(0,1),\ \tilde{T}_1 = 10)$

$\tilde{X}_{s_1}$: 0.79 0.31 0.52 0.16 0.60 0.26 0.65 0.68 0.74 0.45

X_{s_1} : 0 1 0 1 0 1 1 1 0

We have

$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right), \text{ for } m_1 \geq 1$$

Longest increasing run

Scan statistics approach

In the one dimensional problem, let $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ and define $\Pi : \mathbb{R}^2 \to \mathbb{R}$ by

$$\Pi(x,y) = \begin{cases} 1, \text{ if } x < y \\ 0, \text{ otherwise} \end{cases}$$

- the block-factor model becomes: $X_{s_1} = \mathbf{1}_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

Example $(\tilde{X}_{s_1} \sim \mathcal{U}(0,1),\ \tilde{T}_1 = 10)$

$\tilde{X}_{s_1}$:	0.79	0.31	0.52	0.16	0.60	0.26	0.65	0.68	0.74	0.45
X_{s_1} :	0	1	0	1	0	1	1	1	0	

We have

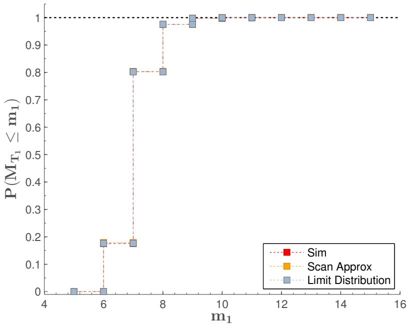
$$\mathbb{P}\left(M_{\tilde{T}_1} \leq m_1\right) = \mathbb{P}\left(L_{T_1} < m_1\right) = \mathbb{P}\left(S_{m_1}(T_1) < m_1\right), \text{ for } m_1 \geq 1$$

Longest increasing run: numerical results

For $\tilde{X}_{s_1} \sim \mathcal{U}([0,1])$, [Novak, 1992] showed that

$$\max_{1 \leq m_1 \leq T_1} \left| \mathbb{P}\left(L_{T_1} < m_1\right) - e^{-T_1 \frac{m_1+1}{(m_1+2)!}} \right| = \mathcal{O}\left(\frac{\ln T_1}{T_1}\right)$$

m_1	Sim	AppH	$E_{total}(1)$	LimApp
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

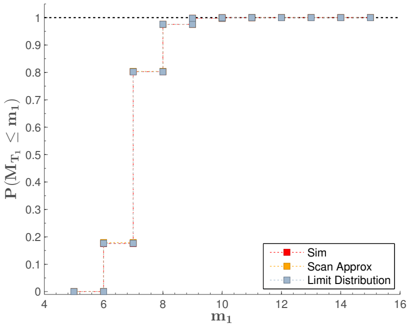


Longest increasing run: numerical results

For $\tilde{X}_{s_1} \sim \mathcal{U}([0,1])$, [Novak, 1992] showed that

$$\max_{1 \le m_1 \le T_1} \left| \mathbb{P}\left(L_{T_1} < m_1\right) - e^{-T_1 \frac{m_1+1}{(m_1+2)!}} \right| = \mathcal{O}\left(\frac{\ln T_1}{T_1}\right)$$

m_1	Sim	AppH	$E_{total}(1)$	LimApp
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



Longest increasing run: numerical results

For $\tilde{X}_{s_1} \sim Geom(p)$, [Louchard and Prodinger, 2003] showed that

$$\mathbb{P}\left(M_{T_1} \leq m_1\right) \sim \exp\left(-\exp\eta\right),$$

$$\eta = \frac{m_1(m_1+1)}{2}\log\frac{1}{1-p} + m_1\log\frac{1}{p} - \log T_1 - \log p + \log D(m_1),$$

$$D(m_1) = \prod_{k=1}^{m_1}\left[1-(1-p)^k\right]\left[1-(1-p)^{m_1+2}\right]$$

m_1	Sim	AppH	$E_{total}(1)$	LimApp
6	0.56445934	0.56997462	0.00255592	0.56810748
7	0.95295406	0.95325180	0.00018554	0.95294598
8	0.99658057	0.99659071	0.00001214	0.99657969
9	0.99979460	0.99979550	0.00000068	0.99979435
10	0.99998950	0.99998950	0.00000003	0.99998947

We used $T_1 = 10000$, $p = 0.1$ and $Iter = 10^5$.

Longest increasing run: numerical results

For $\tilde{X}_{s_1} \sim Geom(p)$, [Louchard and Prodinger, 2003] showed that

$$\mathbb{P}\left(M_{T_1} \leq m_1\right) \sim \exp\left(-\exp \eta\right),$$

$$\eta = \frac{m_1(m_1+1)}{2} \log \frac{1}{1-p} + m_1 \log \frac{1}{p} - \log T_1 - \log p + \log D(m_1),$$

$$D(m_1) = \prod_{k=1}^{m_1} \left[1-(1-p)^k\right] \left[1-(1-p)^{m_1+2}\right]$$

m_1	Sim	AppH	$E_{total}(1)$	LimApp
6	0.56445934	0.56997462	0.00255592	0.56810748
7	0.95295406	0.95325180	0.00018554	0.95294598
8	0.99658057	0.99659071	0.00001214	0.99657969
9	0.99979460	0.99979550	0.00000068	0.99979435
10	0.99998950	0.99998950	0.00000003	0.99998947

We used $T_1 = 10000$, $p = 0.1$ and $Iter = 10^5$.

Outline

Application

Moving average of order q

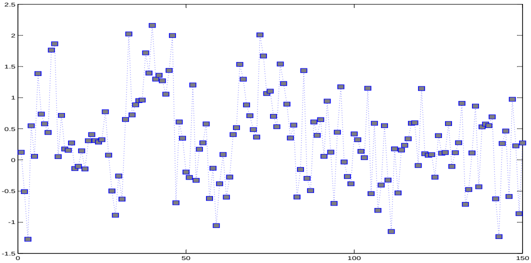
Let $\left(\tilde{X}_n\right)_{n\geq 1}$ be a sequence of i.i.d. $\mathcal{N}(0,\sigma^2)$ r.v.'s.

MA(q)

The sequence $(X_n)_{n\geq 1}$ is said to be an *moving average of order q* (*MA*(*q*)) if

$$X_{s_1} = a_1\tilde{X}_{s_1} + a_2\tilde{X}_{s_1+1} + \cdots + a_{q+1}\tilde{X}_{s_1+q},\ \ s_1 \geq 1,$$

and $(a_1,\ldots,a_{q+1}) \in \mathbb{R}^{q+1}$ not all zero.



Moving average of order q

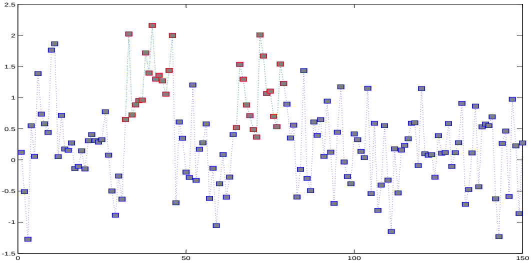
Let $\left(\tilde{X}_n\right)_{n\geq 1}$ be a sequence of i.i.d. $\mathcal{N}(0,\sigma^2)$ r.v.'s.

MA(q)

The sequence $(X_n)_{n\geq 1}$ is said to be an *moving average of order q* (*MA(q)*) if

$$X_{s_1} = a_1\tilde{X}_{s_1} + a_2\tilde{X}_{s_1+1} + \cdots + a_{q+1}\tilde{X}_{s_1+q},\ \ s_1 \geq 1,$$

and $(a_1,\ldots,a_{q+1}) \in \mathbb{R}^{q+1}$ not all zero.



Moving average of order q

Scan statistics approach

Let $d=1$, $x_1^{(1)}=0$, $x_2^{(1)}=q$ thus $c_1=q+1$, $T_1=\tilde{T}_1-q$ and take for $s_1\in\{1,\dots,T_1\}$, the 1-way tensor $\mathcal{X}_{s_1}$

$$\mathcal{X}_{s_1}=\left(\tilde{X}_{s_1},\tilde{X}_{s_1+1},\dots,\tilde{X}_{s_1+q}\right)$$

and define the block-factor $\Pi:\mathbb{R}^{q+1}\to\mathbb{R}$

$$\Pi(x_1,\dots,x_{q+1})=a_1x_1+a_2x_2+\cdots+a_{q+1}x_{q+1}.$$

Example ($MA(2)$)

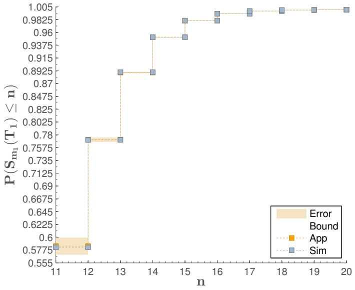
Let $T_1=1000$, $m_1=20$, $\tilde{X}_{s_1}\sim\mathcal{N}(0,1)$ and consider the $MA(2)$

$$X_{s_1}=0.3\tilde{X}_{s_1}+0.1\tilde{X}_{s_1+1}+0.5\tilde{X}_{s_1+2}$$

- Product-type approximation for $MA(2)$: [Wang and Glaz, 2013] and [Wang, 2013]

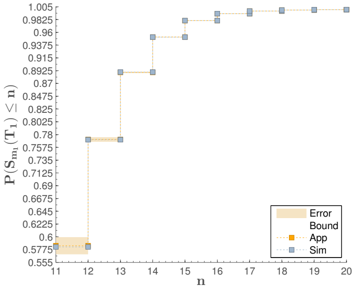
Moving average of order q: numerical results

τ	Sim	AppPT	AppH	$E_{sapp}(1)$	$E_{sf}(1)$	$E_{total}(1)$
11	0.582252	0.589479	0.584355	0.011503	0.003653	0.015156
12	0.770971	0.773700	0.771446	0.002319	0.001691	0.004010
13	0.889986	0.890009	0.889431	0.000434	0.000733	0.001167
14	0.951529	0.954536	0.951723	0.000073	0.000297	0.000370
15	0.980653	0.982433	0.980675	0.000011	0.000113	0.000124
16	0.992827	0.993690	0.992791	0.000001	0.000040	0.000042
17	0.997486	0.995471	0.997499	0.000000	0.000013	0.000014
18	0.999186	0.999411	0.999188	0.000000	0.000004	0.000004
19	0.999754	0.999717	0.999754	0.000000	0.000001	0.000001
20	0.999930	1	0.999930	0.000000	0.000000	0.000000



Moving average of order q: numerical results

τ	Sim	AppPT	AppH	$E_{sapp}(1)$	$E_{sf}(1)$	$E_{total}(1)$
11	0.582252	0.589479	0.584355	0.011503	0.003653	0.015156
12	0.770971	0.773700	0.771446	0.002319	0.001691	0.004010
13	0.889986	0.890009	0.889431	0.000434	0.000733	0.001167
14	0.951529	0.954536	0.951723	0.000073	0.000297	0.000370
15	0.980653	0.982433	0.980675	0.000011	0.000113	0.000124
16	0.992827	0.993690	0.992791	0.000001	0.000040	0.000042
17	0.997486	0.995471	0.997499	0.000000	0.000013	0.000014
18	0.999186	0.999411	0.999188	0.000000	0.000004	0.000004
19	0.999754	0.999717	0.999754	0.000000	0.000001	0.000001
20	0.999930	1	0.999930	0.000000	0.000000	0.000000



Conclusions and perspectives

In this talk:

- introduced the one and two dimensional discrete scan statistics
- introduced a new model of dependence based on block-factor constructions
- presented a unified method for estimating the distribution of the discrete scan statistics both for the i.i.d and the block-factor models
- illustrated an importance sampling algorithm that increases the efficiency of the proposed approximation

Extend and investigate:

- multidimensional continuous scan statistics
- other dependent models
- the influence of the shape of the scanning window
- power of scan statistic based tests under different models

thank you!

Université Lille1
Sciences et Technologies

Amărioarei, A. (2012).
Approximation for the distribution of extremes of one dependent stationary sequences of random variables.
arXiv:1211.5456v1, submitted.

Amărioarei, A. (2014).
Approximations for the multidimensional discrete scan statistics.
PhD thesis, University of Lille 1.

Amărioarei, A. and Preda, C. (2013a).
Approximation for the distribution of three-dimensional discrete scan statistic.
Methodol Comput Appl Probab.

Amărioarei, A. and Preda, C. (2013b).
Approximations for two-dimensional discrete scan statistics in some dependent models.
In *Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013).*

Amărioarei, A. and Preda, C. (2014).
Approximations for two-dimensional discrete scan statistics in some block-factor type dependent models.
Journal of Statistical Planning and Inference, 151-152:107–120.

Boutsikas, M. V. and Koutras, M. V. (2000).
Reliability approximation for Markov chain imbeddable systems.
Methodol. Comput. Appl. Probab., 2:393–411.

Boutsikas, M. V. and Koutras, M. V. (2003).
Bounds for the distribution of two-dimensional binary scan statistics.
Probab. Eng. Inform. Sci., 17:509–525.

Chen, J. and Glaz, J. (1996).
Two-dimensional discrete scan statistics.
Statist. Probab. Lett., 31:59–68.

Chen, J. and Glaz, J. (1997).
Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials.
Advances in the Theory and Practice of Statistics, 1:285–298.

Chen, J. and Glaz, J. (2009).
Scan statistics, chapter 5, Approximations for two-dimensional variable window scan statistics., pages 109–128.
Birkhäuser Boston, Inc., Boston.

Csaki, E. and Foldes, A. (1996).
On the length of theh longest monnotone block.
Studio Scientiarum Mathematicarum Hungarica, 31:35–46.

Devroye, L. (1986).
Non uniform random variate generation.
Springer-Verlag, New York.

Ebneshahrashoob, M. and Sobel, M. (1990).
Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas.
Statist. Probab. Lett., 9:5–11.

Eryilmaz, S. (2006).
A note on runs of geometrically distributed random variables.
Discrete Mathematics, 306:1765–1770.

Fishman, G. (1996).
Monte Carlo: Concepts, Algorithms and Applications.
Springer Series in Operations Research. Springer-Verlag, New York.

Frigessi, A. and Vercellis, C. (1984).

An analysis of Monte Carlo algorithms for counting problems.
Department of Mathematics, University of Milan.

Fu, J. (2001).
Distribution of the scan statistic for a sequence of bistate trials.
J. Appl. Probab., 38:908–916.

Fu, J. C. and Lou, W. (2003).
Distribution theory of runs and patterns and its applications. A finite Markov chain imbedding approach.
World Scientific Publishing Co., Inc., River Edge, NJ.

Gao, T., Ebneshahrashoob, M., and Wu, M. (2005).
An efficient algorithm for exact distribution of discrete scan statistics.
Methodol. Comput. Appl. Probab., 7:1423–1436.

Genz, A. and Bretz, F. (2009).
Computation of Multivariate Normal and T Probabilities.
Springer-Verlag, New York.

Glaz, J. (1990).
A comparison of product-type and Bonferroni-type inequalities in presence of dependence.

In *Symposium on Dependence in Probability and Statistics.*, volume 16 of *IMS Lecture Notes-Monograph Series*, pages 223–235. IMS Lecture Notes.

Glaz, J. and Naus, J. (1991).
Tight bounds and approximations for scan statistic probabilities for discrete data.
Annals of Applied Probability, 1:306–318.

Glaz, J., Naus, J., and Wallenstein, S. (2001).
Scan statistics.
Springer Series in Statistics. Springer-Verlag, New York.

Grabner, P., Knopfmacher, A., and Prodinger, H. (2003).
Combinatorics of geometrically distributed random variables: run statistics.
Theoret. Comput. Sci., 297:261–270.

Grill, K. (1987).
Erdos-Révész type bounds for the length of the longest run from a stationary mixing sequence.
Probab. Theory Relat. Fields, 75:169–179.

Haiman, G. (1999).
First passage time for some stationary processes.
Stochastic Process. Appl., 80:231–248.

Haiman, G. (2000).
Estimating the distributions of scan statistics with high precision.
Extremes, 3:349–361.

Haiman, G. (2007).
Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences.
J. Statist. Plann. Inference, 137:821–828.

Haiman, G. and Preda, C. (2002).
A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process.
Methodol. Comput. Appl. Probab., 4:393–407.

Haiman, G. and Preda, C. (2006).
Estimation for the distribution of two-dimensional discrete scan statistics.
Methodol. Comput. Appl. Probab., 8:373–381.

Karwe, V. and Naus, J. (1997).
New recursive methods for scan statistic probabilities.
Computational Statistics & Data Analysis, 17:389–402.

Kingman, J. (1993).
Poisson processes.
Oxford University Press.

Louchard, G. and Prodinger, H. (2003).
Ascending runs of sequences of geometrically distributed random variables: a probabilistic analysis.
Theoret. Comput. Sci., 304:59–86.

Malley, J., Naiman, D. Q., and Bailey-Wilson, J. (2002).
A compresive method for genome scans.
Human Heredity, 54:174–185.

Naiman, D. Q. and Priebe, C. E. (2001).
Computing scan statistic p values using importance sampling, with applications to genetics and medical image analysis.
J. Comput. Graph. Statist., 10:296–328.

Naiman, D. Q. and Wynn, P. (1997).
Abstract tubes, improved inclusion exclusion identities and inequalities and importance sampling.
The Annals of Statistics, 25:1954–1983.

Naus, J. (1974).
Probabilities for a generalized birthday problem.
Journal of American Statistical Association, 69:810–815.

Naus, J. (1982).
Approximations for distributions of scan statistics.
Journal of American Statistical Association, 77:177–183.

Neil, D. (2006).
Detection of spatial and spatio-temporal clusters.
PhD thesis, School of Computer Science, Carnegie Mellon University.

Neil, D. (2012).
Fast subset scan for spatial pattern detection.
Journal of the Royal Statistical Society, 74(2):337–360.

Novak, S. (1992).
Longest runs in a sequence of m-dependent random variables.
Probab. Theory Relat. Fields, 91:269–281.

Pittel, B. (1981).
Limiting behavior of a process of runs.
Ann. Probab., 9:119–129.

Révész, P. (1983).
Three problems on the llength of increasing runs.
Stochastic Process. Appl., 5:169–179.

Shi, J., Siegmund, D., and Yakir, B. (2007).
Importance sampling for estimating p values in linkage analysis.
Journal of American Statistical Association, 102:929–937.

Wang, X. (2013).
Scan statistics for normal data.
PhD thesis, University of Connecticut.

Wang, X. and Glaz, J. (2013).
A variable window scan statistic for $MA(1)$ process.
In *Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA 2013)*, pages 905–912.

Wang, X., Glaz, J., and Naus, J. (2012).
Approximations and inequalities for moving sums.
Methodol. Comput. Appl. Probab., 14:597–616.

Wu, T.-L. (2013).

On finite Markov chain imbedding technique.

Methodol Comput Appl Probab, 15:453–465.

Product-type approximation and bounds $d = 1$

- Approximation

$$\mathbb{P}\left(S_{m_1}(T_1) \leq \tau\right) \approx Q(2m_1)\left[\frac{Q(3m_1)}{Q(2m_1)}\right]^{\frac{T_1}{m_1}-2},$$

- Lower Bounds

$$\begin{aligned}\mathbb{P}\left(S_{m_1}(T_1) \leq \tau\right) &\leq \frac{Q(2m_1)}{\left[1+\frac{Q(2m_1-1)-Q(2m_1)}{Q(2m_1-1)Q(2m_1)}\right]^{T_1-2m_1}}, \quad T_1 \geq 2m_1\\ &\leq \frac{Q(3m_1)}{\left[1+\frac{Q(2m_1-1)-Q(2m_1)}{Q(3m_1-1)}\right]^{T_1-3m_1}}, \quad T_1 \geq 3m_1\end{aligned}$$

- Upper Bounds

$$\begin{aligned}\mathbb{P}\left(S_{m_1}(T_1) \leq \tau\right) &\leq Q(2m_1)\left[1 - Q(2m_1-1) + Q(2m_1)\right]^{T_1-2m_1}, \quad T_1 \geq 2m_1\\ &\leq Q(3m_1)\left[1 - Q(2m_1-1) + Q(2m_1)\right]^{T_1-3m_1}, \quad T_1 \geq 3m_1\end{aligned}$$

The values $Q(2m_1-1), Q(2m_1), Q(3m_1-1), Q(3m_1)$ are computed using [Karwe and Naus, 1997] algorithm.

Return

Product-type approximation and bounds $d = 2$

- Approximation (Bernoulli)

$$\mathbb{P}\left(S_{m_1,m_2}(T_1, T_2) \leq k\right) \approx \frac{Q(m_1,m_2)^{(T_1-m_1-1)(T_2-m_2-1)}\,Q(m_1+1,m_2+1)^{(T_1-m_1)(T_2-m_2)}}{Q(m_1,m_2+1)^{(T_1-m_1-1)(T_2-m_2)}\,Q(m_1+1,m_2)^{(T_1-m_1)(T_2-m_2-1)}}$$

- Approximation (binomial and Poisson)

$$\mathbb{P}\left(S_{m_1,m_2}(T_1, T_2) \leq k\right) \approx \frac{Q(m_1+1,m_2+1)^{(T_1-m_1)(T_2-m_2)}}{Q(m_1+1,m_2)^{(T_1-m_1)(T_2-m_2-1)}} \times \frac{Q(m_1,2m_2-1)^{(T_1-m_1-1)(T_2-2m_2)}}{Q(m_1,2m_2)^{(T_1-m_1-1)(T_2-2m_2+1)}}$$

To compute the unknown variables we use

- $Q(m_1, 2m_2 - 1)$ and $Q(m_1, 2m_2)$ - adaptation of [Karwe and Naus, 1997] algorithm
- $Q(m_1 + 1, m_2)$ and $Q(m_1 + 1, m_2 + 1)$ - conditioning

Return

Approach

[Fu, 2001] applied the Markov Chain Imbedding Technique to find the distribution of binary scan statistics.

Main Idea

Express the distribution of the $S_{m_1}(T_1)$ in terms of the waiting time distribution of a special compound pattern

- define for $0 \leq k \leq m_1$

$$\mathcal{F}_{m_1,k} = \{\Lambda_i | \Lambda_1 = \underbrace{1\dots1}_{k}, \Lambda_2 = 10\underbrace{1\dots1}_{k-1}, \dots, \Lambda_l = \overbrace{\underbrace{1\dots1}_{k-1}0\dots01}^{m_1}\}$$

$$|\mathcal{F}_{m_1,k}| = \sum_{j=0}^{m_1-k_1} \binom{k-2+j}{j}$$

- the compound pattern $\Lambda = \cup_{i=1}^{l}\Lambda_i,\ \Lambda_i \in \mathcal{F}_{m_1,k}$

$$\mathbb{P}(S_{m_1}(T_1) < k) = \mathbb{P}(W(\Lambda) \geq T_1 + 1).$$

$$\mathbb{P}(S_{m_1}(T_1) < k) = \boldsymbol{\xi} \boldsymbol{N}^{T_1}\mathbf{1}^\intercal, \text{where } \boldsymbol{\xi} = (1, 0, \dots, 0)$$

Université Lille1

Return

Example

Consider the i.i.d. two-state sequence $(X_i)_{i\in\{1,2,\dots,T_1\}}$ with $p=\mathbb{P}(X_1=1)$ and $q=\mathbb{P}(X_1=0)$.

- A realisation for $T_1=20$

  00101011101101010110

- For $k=3$ and $m_1=4$

  $\mathcal{F}_{4,3}=\{\Lambda_1=111,\Lambda_2=1011,\Lambda_3=1101\}$

- The state space

  $\Omega=\{\emptyset,0,1,10,11,101,110,\alpha_1,\alpha_2,\alpha_3\}$

- the principal matrix:

$$N=\begin{pmatrix} 0 & q & p & 0 & 0 & 0 & 0\\ 0 & q & p & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & q & p & 0 & 0\\ 0 & q & 0 & 0 & 0 & p & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & q\\ 0 & 0 & 0 & q & 0 & 0 & 0\\ 0 & q & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Return

Selected Values for $K(\cdot)$ and $\Gamma(\cdot)$

Table 5 : Selected values for $K(\cdot)$ and $\Gamma(\cdot)$

$1 - q_1$	$K(1 - q_1)$	$\Gamma(1 - q_1)$
0.1	38.63	480.69
0.05	21.28	180.53
0.025	17.56	145.20
0.01	15.92	131.43

Return

Selected Values for $K(\cdot)$ and $\Gamma(\cdot)$

Table 5 : Selected values for $K(\cdot)$ and $\Gamma(\cdot)$

$1-q_1$	$K(1-q_1)$	$\Gamma(1-q_1)$
0.1	38.63	480.69
0.05	21.28	180.53
0.025	17.56	145.20
0.01	15.92	131.43

Return

Error bounds: approximation error

Approximation error

$$E_{app}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \sum_{t_1,\dots,t_{s-1}\in\{2,3\}} F_{t_1,\dots,t_{s-1}} \left(1 - \gamma_{t_1,\dots,t_{s-1},2} + B_{t_1,\dots,t_{s-1},2}\right)^2,$$

where for $2 \leq s \leq d$

$$F_{t_1,\dots,t_{s-1}} = F\left(Q_{t_1,\dots,t_{s-1},2}, L_s - 1\right), \; F = F\left(Q_2, L_1 - 1\right),$$

$$B_{t_1\dots,t_{s-1}} = (L_s - 1)\left[F_{t_1,\dots,t_{s-1}} \left(1 - \gamma_{t_1,\dots,t_{s-1},2} + B_{t_1\dots,t_{s-1},2}\right)^2 + \sum_{t_s\in\{2,3\}} B_{t_1\dots,t_s}\right],$$

$$B_{t_1\dots,t_{d-1}} = (L_d - 1)F_{t_1,\dots,t_{d-1}} \left(1 - \gamma_{t_1,\dots,t_{d-1},2} + B_{t_1\dots,t_{d-1},2}\right)^2, \; B_{t_1\dots,t_d} = 0,$$

and for $s = 1$: $\sum_{t_1,t_0\in\{2,3\}} x = x$, $F_{t_1,t_0} = F$, $\gamma_{t_1,t_0,2} = \gamma_2$ and $B_{t_1,t_0,2} = B_2$.

Return

Error bounds: simulation errors

Simulation errors

$$E_{sf}(d) = (L_1 - 1)\dots(L_d - 1) \sum_{t_1,\dots,t_d \in \{2,3\}} \beta_{t_1,\dots,t_d}$$

$$E_{sapp}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \sum_{t_1,\dots,t_{s-1} \in \{2,3\}} F_{t_1,\dots,t_{s-1}} \Big(1 - \hat{Q}_{t_1,\dots,t_{s-1},2}$$
$$+A_{t_1,\dots,t_{s-1},2} + C_{t_1,\dots,t_{s-1},2}\Big)^2$$

where for $2 \leq s \leq d$

$$A_{t_1,\dots,t_{s-1}} = (L_s - 1)\dots(L_d - 1) \sum_{t_s,\dots,t_d \in \{2,3\}} \beta_{t_1,\dots,t_d},\ A_{t_1,\dots,t_d} = \beta_{t_1,\dots,t_d}$$

$$C_{t_1\dots,t_{s-1}} = (L_s - 1)\Big[F_{t_1,\dots,t_{s-1}}\Big(1 - \hat{Q}_{t_1,\dots,t_{s-1},2} + A_{t_1\dots,t_{s-1},2} + C_{t_1\dots,t_{s-1},2}\Big)^2$$
$$+ \sum_{t_s \in \{2,3\}} C_{t_1\dots,t_s}\Big]$$

Return

Université Lille1 Sciences et Technologies

Discrete scan statistics for normal data

Consider $d = 1$ and let $2 \leq m_1 \leq T_1$, m_1 and T_1 be positive integers

- $X_{s_1} \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d., $1 \leq s_1 \leq T_1$

The variables $Y_{i_1} = \sum_{s_1=i_1}^{i_1+m_1-1} X_{s_1}$ follow a multivariate normal distribution with mean $\bar{\mu} = m_1\mu$ and covariance matrix $\Sigma = (\Sigma_{i_1,j_1})$

$$\Sigma_{i_1,j_1} = Cov\left[Y_{i_1}, Y_{j_1}\right] = \begin{cases} (m_1 - |i_1 - j_1|)\,\sigma^2 & , |i_1 - j_1| < m_1 \\ 0 & , \text{otherwise.} \end{cases}$$

Return

Step 2 in Algorithm 2

Step 2 requires to sample:

- $Y_{i_1^{(k)}}$ from the tail distribution $\mathbb{P}\left(Y_{i_1^{(k)}} \geq \tau\right)$ ([Devroye, 1986])
- for the other indices, from the conditional distribution given $\left\{Y_{i_1^{(k)}} \geq \tau\right\}$

For $\mathbf{W}_1 = \left(Y_1, \dots, Y_{i_1^{(k)}-1}\right)$ and $\mathbf{W}_2 = \left(Y_{i_1^{(k)}+1}, \dots, Y_{T_1-m_1+1}\right)$

$$\overline{\mathbf{W}}_1 = \mathbf{W}_1|(Y_{i_1^{(k)}} = t) \sim \mathcal{N}\left(\mu_{w_1|t}, \Sigma_{w_1|t}\right) \text{ and } \overline{\mathbf{W}}_2 = \mathbf{W}_2|(Y_{i_1^{(k)}} = t) \sim \mathcal{N}\left(\mu_{w_2|t}, \Sigma_{w_2|t}\right)$$

where for $i \in \{1,2\}$,

$$\mu_{w_i|t} = \mathbb{E}[\mathbf{W}_i] + \frac{1}{Var[Y_{i_1^{(k)}}]} Cov[\mathbf{W}_i, Y_{i_1^{(k)}}](t - \mathbb{E}[Y_{i_1^{(k)}}]),$$

$$\Sigma_{w_i|t} = Cov(\mathbf{W}_i) - \frac{1}{Var[Y_{i_1^{(k)}}]} Cov[\mathbf{W}_i, Y_{i_1^{(k)}}] Cov^T[\mathbf{W}_i, Y_{i_1^{(k)}}].$$

Return

Cumulative counts method

Idea

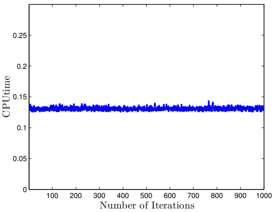
Precompute a matrix of cumulative counts $\boldsymbol{M}$ using dynamic programming and express the variables of interest as differences.

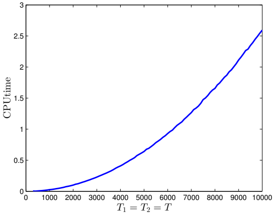
- efficiently searches for the locality statistics over $\mathcal{R}_d$ in constant time

Example ($d = 2,\ T_1 = T_2 = T,\ m_1 = m_2 = m$)

The matrix M has the entries $M(i,j) = \sum_{k=1}^{i}\sum_{l=1}^{j} X_{k,l}$, so the locality statistic is

$$Y_{i_1,i_2} = M(i_1+m-1, i_2+m-1) - M(i_1+m-1, i_2-1) - M(i_1-1, i_2+m-1) + M(i_1-1, i_2-1)$$





Return

Alternative approaches

Several other methods were proposed:

I) [Genz and Bretz, 2009] developed a quasi Monte Carlo algorithm for numerically approximate the distribution of a multivariate normal, the algorithm was implemented in R and Matlab ([Wang and Glaz, 2013], [Wang, 2013])

II) [Shi et al., 2007] introduced another IS algorithm (Algo 3)

- idea: imbed the probability measure under H_0 into an exponential family

Details Algo 3

To measure the efficiency of the methods we evaluate the *relative efficiency* introduced by [Malley et al., 2002]

$$Rel\ Eff = \frac{\sigma^2_{method\ 1} \times CPU\ Time_{method\ 1}}{\sigma^2_{method\ 2} \times CPU\ Time_{method\ 2}}$$

Return

IS algorithm [Shi et al., 2007]

Algorithm 3 Second Importance Sampling Algorithm for Scan Statistics

Take $d\mathbb{P}_{\xi,r_1} = \frac{e^{\xi Y_{r_1}}}{\mathbb{E}_{H_0}\left[e^{\xi Y_{r_1}}\right]} d\mathbb{P}_{H_0}$ and compute

$$\xi \approx \frac{\tau}{m_1\sigma^2} - \frac{\mu}{\sigma^2}, \quad \mathbb{E}_{\xi,r_1}\left[Y_{i_1}\right] = \xi Cov_{H_0}\left[Y_{i_1}, Y_{r_1}\right] + m_1\mu, \quad Cov_{\xi,r_1}\left[Y_{i_1}, Y_{j_1}\right] = Cov_{H_0}\left[Y_{i_1}, Y_{j_1}\right]$$

Repeat for each k from 1 to $ITER$ (iterations number)

1: Generate uniformly $i_1^{(k)}$ from the set $\{1, \ldots, T_1 - m_1 + 1\}$.

2: Given $i_1^{(k)}$, generate the Gaussian process Y_{i_1} according to the new measure $d\mathbb{P}_{\xi, i_1^{(k)}}$.

3: Compute $\widehat{\rho}_k(1)$ based on

$$\widehat{\rho}_k(1) = \frac{T_1 - m_1 + 1}{\sum_{j_1=1}^{T_1-m_1+1} e^{\xi Y_{j_1} - m_1\left(\mu\xi + \frac{\sigma^2\xi^2}{2}\right)}} 1_{\left\{S_{m_1}(T_1) \geq \tau\right\}}$$

End Repeat

Return

$$\widehat{\rho}(1) = \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(1), \quad Var\left[\widehat{\rho}(1)\right] \approx \frac{1}{ITER - 1}\sum_{k=1}^{ITER}\left(\widehat{\rho}_k(1) - \frac{1}{ITER}\sum_{k=1}^{ITER}\widehat{\rho}_k(1)\right)^2$$

Return

Numerical results

All the results are compared with respect to Algo 2 for $ITER = 10000$

Table 6 : Algorithm [Genz and Bretz, 2009], IS (Algo 2) and the relative efficiency (Rel Eff)

T_1	m_1	τ	Genz	Err Genz	IS Algo 2	Err Algo 2	Rel Eff
200	15	12	0.932483	0.000732	0.933215	0.000743	7
500	25	18	0.976117	0.000460	0.975797	0.000425	518
750	30	24	0.998454	0.000125	0.998493	0.000024	688
800	40	30	0.999752	0.000029	0.999742	0.000004	617

Table 7 : Naive Monte Carlo (MC), IS (Algo 2) and the relative efficiency (Rel Eff)

T_1	m_1	τ	MC	Err MC	IS Algo 2	Err Algo 2	Rel Eff
200	15	12	0.932624	0.000694	0.933215	0.000743	15
500	25	18	0.975880	0.000425	0.975797	0.000425	33
750	30	24	0.998515	0.000061	0.998493	0.000024	101
800	40	30	0.999741	0.000009	0.999742	0.000004	602

Université Lille1 Sciences et Technologies

Return

Numerical results

Table 8 : IS algorithms (Algo 2 and Algo 2) and the relative efficiency (Rel Eff)

T_1	m_1	τ	IS Algo 2	Err Algo 2	IS Algo 2	Err Algo 2	Rel Eff
200	15	12	0.932744	0.000839	0.933215	0.000743	3
500	25	18	0.976105	0.000448	0.975797	0.000425	3.5
750	30	24	0.998508	0.000032	0.998493	0.000024	3.5
800	40	30	0.999740	0.000006	0.999742	0.000004	3.6

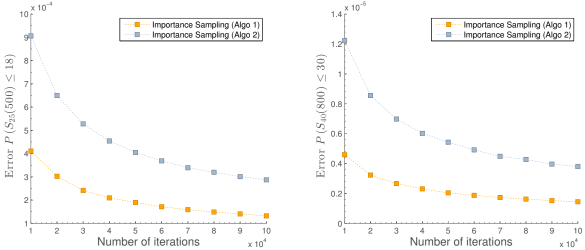


Figure 1 : The evolution of simulation error in IS Algorithm 2 and IS Algorithm 2

Return

Numerical results for big scanning window

Table 9 : $n = 1, p = 0.01, m_1 = 10^4, T_1 = 10^6, It_{App} = 10^4$

k	AppH	EsappH	AppHIS	EtotalIS	AppPT	LowB	UppB
135	0.709261	0.001763	0.664332	0.169866	0.709116	0.708348	0.709312
136	0.773735	0.000956	0.772472	0.050303	0.773652	0.773187	0.773769
137	0.826917	0.000513	0.831974	0.031697	0.826872	0.826599	0.826939
138	0.869618	0.000272	0.869167	0.034322	0.869593	0.869439	0.869631
139	0.903125	0.000142	0.905000	0.019600	0.903112	0.903027	0.903133
140	0.928908	0.000073	0.928772	0.013337	0.928901	0.928855	0.928912
141	0.948413	0.000037	0.949536	0.010616	0.948410	0.948386	0.948415
142	0.962952	0.000019	0.962711	0.013716	0.962951	0.962938	0.962953
143	0.973649	0.000009	0.971999	0.004796	0.973648	0.973641	0.973649
144	0.981425	0.000005	0.981516	0.003736	0.981425	0.981422	0.981425
145	0.987019	0.000002	0.987272	0.001966	0.987018	0.987017	0.987019
146	0.991002	0.000001	0.991091	0.001980	0.991002	0.991001	0.991002
147	0.993812	0.000000	0.993717	0.001308	0.993812	0.993811	0.993812
148	0.995777	0.000000	0.995720	0.000767	0.995777	0.995777	0.995777
149	0.997140	0.000000	0.996961	0.000478	0.997140	0.997140	0.997140
150	0.998077	0.000000	0.998072	0.000587	0.998077	0.998077	0.998077
151	0.998716	0.000000	0.998767	0.000228	0.998716	0.998716	0.998716
152	0.999149	0.000000	0.999097	0.000213	0.999149	0.999149	0.999149
153	0.999440	0.000000	0.999445	0.000096	0.999440	0.999440	0.999440
154	0.999634	0.000000	0.999638	0.000096	0.999634	0.999634	0.999634
155	0.999762	0.000000	0.999758	0.000045	0.999762	0.999762	0.999762
156	0.999847	0.000000	0.999855	0.000032	0.999847	0.999847	0.999847
157	0.999902	0.000000	0.999903	0.000019	0.999902	0.999902	0.999902
158	0.999938	0.000000	0.999939	0.000011	0.999938	0.999938	0.999938
159	0.999961	0.000000	0.999954	0.000012	0.999961	0.999961	0.999961

Return

Numerical results for big scanning window

Table 9 : $n = 1, p = 0.01, m_1 = 10^4, T_1 = 10^6, It_{App} = 10^4$

k	AppH	EsappH	AppHIS	EtotalIS	AppPT	LowB	UppB
135	0.709261	0.001763	0.664332	0.169866	0.709116	0.708348	0.709312
136	0.773735	0.000956	0.772472	0.050303	0.773652	0.773187	0.773769
137	0.826917	0.000513	0.831974	0.031697	0.826872	0.826599	0.826939
138	0.869618	0.000272	0.869167	0.034322	0.869593	0.869439	0.869631
139	0.903125	0.000142	0.905000	0.019600	0.903112	0.903027	0.903133
140	0.928908	0.000073	0.928772	0.013337	0.928901	0.928855	0.928912
141	0.948413	0.000037	0.949536	0.010616	0.948410	0.948386	0.948415
142	0.962952	0.000019	0.962711	0.013716	0.962951	0.962938	0.962953
143	0.973649	0.000009	0.971999	0.004796	0.973648	0.973641	0.973649
144	0.981425	0.000005	0.981516	0.003736	0.981425	0.981422	0.981425
145	0.987019	0.000002	0.987272	0.001966	0.987018	0.987017	0.987019
146	0.991002	0.000001	0.991091	0.001980	0.991002	0.991001	0.991002
147	0.993812	0.000000	0.993717	0.001308	0.993812	0.993811	0.993812
148	0.995777	0.000000	0.995720	0.000767	0.995777	0.995777	0.995777
149	0.997140	0.000000	0.996961	0.000478	0.997140	0.997140	0.997140
150	0.998077	0.000000	0.998072	0.000587	0.998077	0.998077	0.998077
151	0.998716	0.000000	0.998767	0.000228	0.998716	0.998716	0.998716
152	0.999149	0.000000	0.999097	0.000213	0.999149	0.999149	0.999149
153	0.999440	0.000000	0.999445	0.000096	0.999440	0.999440	0.999440
154	0.999634	0.000000	0.999638	0.000096	0.999634	0.999634	0.999634
155	0.999762	0.000000	0.999758	0.000045	0.999762	0.999762	0.999762
156	0.999847	0.000000	0.999855	0.000032	0.999847	0.999847	0.999847
157	0.999902	0.000000	0.999903	0.000019	0.999902	0.999902	0.999902
158	0.999938	0.000000	0.999939	0.000011	0.999938	0.999938	0.999938
159	0.999961	0.000000	0.999954	0.000012	0.999961	0.999961	0.999961

Université Lille1

Return

Error bounds: approximation error

Approximation error

$$E_{app}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \sum_{t_1,\dots,t_{s-1}\in\{2,3\}} F_{t_1,\dots,t_{s-1}} \left(1 - \gamma_{t_1,\dots,t_{s-1},2} + B_{t_1,\dots,t_{s-1},2}\right)^2,$$

where for $2 \leq s \leq d$

$$F_{t_1,\dots,t_{s-1}} = F\left(Q_{t_1,\dots,t_{s-1},2}, L_s - 1\right), \; F = F\left(Q_2, L_1 - 1\right),$$

$$B_{t_1\dots,t_{s-1}} = (L_s - 1)\left[F_{t_1,\dots,t_{s-1}}\left(1 - \gamma_{t_1,\dots,t_{s-1},2} + B_{t_1\dots,t_{s-1},2}\right)^2 + \sum_{t_s\in\{2,3\}} B_{t_1\dots,t_s}\right],$$

$$B_{t_1\dots,t_{d-1}} = (L_d - 1)F_{t_1,\dots,t_{d-1}}\left(1 - \gamma_{t_1,\dots,t_{d-1},2} + B_{t_1\dots,t_{d-1},2}\right)^2, \; B_{t_1\dots,t_d} = 0,$$

and for $s = 1$: $\sum_{t_1,t_0\in\{2,3\}} x = x$, $F_{t_1,t_0} = F$, $\gamma_{t_1,t_0,2} = \gamma_2$ and $B_{t_1,t_0,2} = B_2$.

Return

Error bounds: simulation errors

Simulation errors

$$E_{sf}(d) = (L_1 - 1)\dots(L_d - 1) \sum_{t_1,\dots,t_d \in \{2,3\}} \beta_{t_1,\dots,t_d}$$

$$E_{sapp}(d) = \sum_{s=1}^{d} (L_1 - 1)\cdots(L_s - 1) \sum_{t_1,\dots,t_{s-1} \in \{2,3\}} F_{t_1,\dots,t_{s-1}} \Big(1 - \hat{Q}_{t_1,\dots,t_{s-1},2} + A_{t_1,\dots,t_{s-1},2} + C_{t_1,\dots,t_{s-1},2}\Big)^2$$

where for $2 \leq s \leq d$

$$A_{t_1,\dots,t_{s-1}} = (L_s - 1)\dots(L_d - 1) \sum_{t_s,\dots,t_d \in \{2,3\}} \beta_{t_1,\dots,t_d},\ A_{t_1,\dots,t_d} = \beta_{t_1,\dots,t_d}$$

$$C_{t_1\dots,t_{s-1}} = (L_s - 1)\Big[F_{t_1,\dots,t_{s-1}} \Big(1 - \hat{Q}_{t_1,\dots,t_{s-1},2} + A_{t_1\dots,t_{s-1},2} + C_{t_1\dots,t_{s-1},2}\Big)^2 + \sum_{t_s \in \{2,3\}} C_{t_1\dots,t_s}\Big]$$

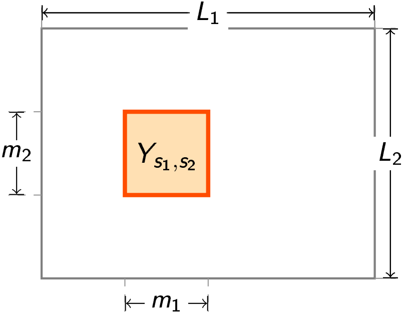
Return

Université Lille1 Sciences et Technologies

One and two dimensional continuous scan statistics

Scan statistics associated to a Poisson process

Let N be a two (one) dimensional Poisson process of intensity λ and $m_j \leq T_j$, $1 \leq j \leq 2$ be positive integers



- Define for $0 \leq s_j \leq T_j - m_j$,

$$Y_{s_1,s_2} = N\left([s_1, s_1 + m_1) \times [s_2, s_2 + m_2)\right)$$

- The two dimensional scan statistic,

$$S_{m_1,m_2}(\lambda, T_1, T_2) = \max_{\substack{0 \leq s_1 \leq T_1 - m_1 \\ 0 \leq s_2 \leq T_2 - m_2}} Y_{s_1,s_2}$$

Observe that by applying the mapping theorem ([Kingman, 1993]) we have

$$\mathbb{P}\left(S_{m_1,m_2}(\lambda, T_1, T_2) \leq \tau\right) = \mathbb{P}\left(S_{1,1}\left(\lambda m_1 m_2, \frac{T_1}{m_1}, \frac{T_2}{m_2}\right) \leq \tau\right)$$

Approximation process in two dimensions