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

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
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
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

\[
Y_{i_1} = \sum_{j=i_1}^{i_1+m_1-1} X_j
\]

then the discrete one dimensional scan statistics is defined as

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

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

\[
\begin{array}{cccccccccccccccccc}
0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0
\end{array}
\]

\( Y_{i_1} = 1 \)
**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 $m_1$
  \[
  \mathbb{P}(W_{m_1,k} \leq T_1) = \mathbb{P}(S_{m_1}(T_1) \geq k)
  \]

- $D_{T_1}(k)$ - the length of the smallest window that contains at least $k$ successes
  \[
  \mathbb{P}(D_{T_1}(k) \leq m_1) = \mathbb{P}(S_{m_1}(T_1) \geq k)
  \]

- $L_{T_1}$ - the length of the longest success run
  \[
  \mathbb{P}(L_{T_1} \geq m_1) = \mathbb{P}(S_{m_1}(T_1) \geq m_1) = \mathbb{P}(S_{m_1}(T_1) = m_1)
  \]
**Problem and Approaches**

**Problem**

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

\[
P(S_{m_1}(T_1) \leq \tau).
\]

**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

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
Two dimensional discrete scan statistics
Introducing the Model

Let $T_1, T_2$ be positive integers

- Rectangular region
  $R_2 = [0, T_1] \times [0, T_2]$

- $X_{s_1, s_2}$ i.i.d. integer r.v.'s
  $1 \leq s_1 \leq T_1$
  $1 \leq s_2 \leq T_2$
  - Bernoulli($B(1, p)$)
  - Binomial($B(n, p)$)
  - Poisson($P(\lambda)$)

- $X_{s_1, s_2}$ 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

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

The two dimensional scan statistic,

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

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. \( B(p) \)

\( H_1: \) 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 \( X_{s_1, s_2} \sim B(p') \), \( p' > p \) and \( X_{s_1, s_2} \sim B(p) \) 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

\[ Q_m(T) = P(S_m(T) \leq \tau) \]

with \( m = (m_1, m_2) \) and \( T = (T_1, T_2) \)

**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]
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 \(\sigma\)-fields generated by \(\{Z_1, \ldots, Z_h\}\) and \(\{Z_{h+m+1}, \ldots\}\) 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 \(t_1, \ldots, t_k\) the families \(\{Z_{t_1}, \ldots, Z_{t_k}\}\) and \(\{Z_{t_1+h}, \ldots, Z_{t_k+h}\}\) have the same joint distribution.

**Notation**

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

\[
q_n = q_n(x) = \mathbb{P}(\max(Z_1, \ldots, Z_n) \leq x)
\]
**The main result**

**Theorem [Haiman, 1999]**

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

\[
|q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n}| \leq n\Delta^H_2 (1 - q_1)^2
\]

\( \Delta^H_2 = 3.3 + \frac{9}{n} + \left[ 15.51n(1 - q_1) + \frac{561}{n} \right] (1 - q_1) . \)

**Theorem [Amārioarei, 2012]**

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

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

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

- Increased range of applicability
- Sharper error bounds
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

\[
\left| q_n - \frac{2q_1 - q_2}{[1 + q_1 - q_2 + 2(q_1 - q_2)^2]^n} \right| \leq n\Delta^H_2 (1 - q_1)^2
\]

- \( \Delta^H_2 = 3.3 + \frac{9}{n} + \left[ 15.51n(1 - q_1) + \frac{561}{n} \right] (1 - q_1) \).

**Theorem [Amarioarei, 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
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]
Scan Statistics (i.i.d model)  Scan statistics and 1-dependent sequences

**$S_m(T)$ 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

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

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

- Observe

$$S_m(T) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}$$

**Example (One dimensional case)**

$$X_1, X_2, \ldots, X_{m_1 - 1}, X_{m_1}, \ldots, X_{2(m_1 - 1)}, X_{2m_1 - 1}, \ldots, X_{3(m_1 - 1)}, X_{3m_1 - 2}, \ldots, X_{4(m_1 - 1)}$$
**Scan Statistics (i.i.d model)**

**Scan statistics and 1-dependent sequences**

\[ S_m(T) \text{ viewed as maximum of 1-dependent r.v.'s} \]

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

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

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

- Observe

\[
S_m(T) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}
\]

**Example (One dimensional case)**

\[
X_1, X_2, \ldots, X_{m_1 - 1}, X_{m_1}, \ldots, X_{2(m_1 - 1)}, X_{2m_1 - 1}, \ldots, X_{3(m_1 - 1)}, X_{3m_1 - 2}, \ldots, X_{4(m_1 - 1)}
\]
Scan Statistics (i.i.d model)

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

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

- Observe
  \[
  S_m(T) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}
  \]

Example (One dimensional case)

$X_1, X_2, \ldots, X_{m_1 - 1}, X_{m_1}, \ldots, X_{2(m_1 - 1)}, X_{2m_1 - 1}, \ldots, X_{3(m_1 - 1)}, X_{3m_1 - 2}, \ldots, X_{4(m_1 - 1)}$
$S_m(T)$ 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
  \[
  Z_{k_1} = \max_{(k_1-1)(m_1-1)+1 \leq i_1 \leq k_1(m_1-1)} \max_{1 \leq i_2 \leq (L_2-1)(m_2-1)} Y_{i_1,i_2}
  \]

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

- Observe
  \[
  S_m(T) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}
  \]

Example (One dimensional case)

$X_1, X_2, \ldots, X_{m_1-1}, X_{m_1}, \ldots, X_{2(m_1-1)}, X_{2m_1-1}, \ldots, X_{3(m_1-1)}, X_{3m_1-2}, \ldots, X_{4(m_1-1)}$
$S_m(T)$ viewed as maximum of 1-dependent r.v.'s

Example (Two dimensional case)
**Approximation process: First step**

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

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

If $1 - Q_2 \leq 0.1$ then

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

**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 $S_m(T)$ is an iterative process. The second step becomes:

- Define for $t_1 \in \{2, 3\}$ and $k_2 \in \{1, 2, \ldots, L_2 - 1\}$
  \[ Z_{k_2}^{(t_1)} = \max_{1 \leq i_1 \leq (t_1-1)(m_1-1)} Y_{i_1,i_2} \]

- \( \{Z_1^{(t_1)}, \ldots, Z_{L_2-1}^{(t_1)}\} \) 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
  \[ |Q_{t_1} - H(Q_{t_1,2}, Q_{t_1,3}, L_2)| \leq (L_2 - 1)F(Q_{t_1,2}, L_2 - 1)(1 - Q_{t_1,2})^2 \]
ILLUSTRATION FOR THE TWO DIMENSIONAL CASE

Find Approximation

A. Amãrioarei (INRIA)
ILLUSTRATION FOR THE TWO DIMENSIONAL CASE

First Step Approximation
ILLUSTRATION FOR THE TWO DIMENSIONAL CASE

Second Step
Approximation

A. Amarioarei (INRIA) 1d and 2d Scan Statistics SSB Seminar 26 / 62
**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

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

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

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

**OBJECTIVE**

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

We observe that

$$|Q_m(T) - H(\hat{Q}_2, \hat{Q}_3, L_1)| \leq |Q_m(T) - H(\gamma_2, \gamma_3, L_1)| + |H(\gamma_2, \gamma_3, L_1) - H(\hat{Q}_2, \hat{Q}_3, L_1)|$$

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

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

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

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

**OBJECTIVE**

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

We observe that

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

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

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

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

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

**OBJECTIVE**

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

We observe that

$$|Q_m(T) - H(\hat{Q}_2, \hat{Q}_3, L_1)| \leq |Q_m(T) - H(\gamma_2, \gamma_3, L_1)| + |H(\gamma_2, \gamma_3, L_1) - H(\hat{Q}_2, \hat{Q}_3, L_1)|$$

$$E_{app}(2) \leq E_{sapp}(2)$$

$$E_{sf}(2)$$

The quantities $\hat{Q}_{t_1,t_2}$ will be estimated by Monte Carlo simulations.
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
Simulation methods and computational aspects
Scan Statistics (i.i.d model) Simulation methods and computational aspects

Naive Hit-or-Miss MC

Objective

Find an estimate for $P_{H_0}(S_m(T) \geq \tau)$.

Algorithm 1 Classical Monte Carlo algorithm for scan statistics

Begin
Repeat for each $k$ from 1 to $ITER$ (iterations number)
1: Generate $X^{(k)} = \{X_{s_1,s_2}^{(k)}, 1 \leq s_j \leq T_j, 1 \leq j \leq 2\}$ under $H_0$
2: Compute the two dimensional scan statistics $S_m^{(k)}(T)$ over $X^{(k)}$
End Repeat
Return

$\hat{p}_{MC} = \frac{1}{ITER} \sum_{i=1}^{ITER} 1\{S_m^{(i)}(T) \geq \tau\}$, $s.e. MC = \sqrt{\frac{\hat{p}_{MC}(1 - \hat{p}_{MC})}{ITER}}$

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

\[ P_{H_0}(S_m(T) \geq \tau) = \mathbb{P}\left( \bigcup_{i_1=1}^{T_1-m_1+1} \bigcup_{i_2=1}^{T_2-m_2+1} E_{i_1,i_2} \right) = \int G(x)f(x) \, dx \]

where \( E_{i_1,i_2} = \{ Y_{i_1,i_2} \geq \tau \} \), \( G(x) = 1_{E}(x) \), \( E = \bigcup_{i_1=1}^{T_1-m_1+1} \bigcup_{i_2=1}^{T_2-m_2+1} E_{i_1,i_2} \) 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

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

and we observe that

\[ P_{H_0}(S_m(T) \geq \tau) = B(2) \rho(2) \]

- the Bonferroni upper bound \( B(2) \)
- the correction factor \( \rho(2) \) between 0 and 1

\[ \rho(2) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} p_{i_1,i_2} \int \frac{1}{C(Y)} dP_{H_0}(\cdot \mid E_{i_1,i_2}) \]

where

\[ p_{i_1,i_2} = \frac{1}{(T_1-m_1+1)(T_2-m_2+1)} \]

\[ C(Y) = \sum_{i_1=1}^{T_1-m_1+1} \sum_{i_2=1}^{T_2-m_2+1} 1_{E_{i_1,i_2}} \]
**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, \ldots, $T_1 - m_1 + 1$\} $\times$ \{1, \ldots, $T_2 - m_2 + 1$\}. 

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

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

**End Repeat**

**Return**

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

**End**
**Implementation problems**

Algorithm 2 presents two main difficulties:

A) being able to sample from the conditional distribution of $X$ 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

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
Numerical examples
ONE DIMENSIONAL CASE: $X_i \sim B(n, p)$

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

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>Exact</th>
<th>Glaz et al.</th>
<th>Our Approximation</th>
<th>Approximation Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.810209</td>
<td>0.810216</td>
<td>0.810404</td>
<td>0.001111</td>
<td>0.809903</td>
<td>0.810439</td>
</tr>
<tr>
<td>2</td>
<td>0.995764</td>
<td>0.995764</td>
<td>0.995764</td>
<td>$3 \times 10^{-7}$</td>
<td>0.995764</td>
<td>0.995764</td>
</tr>
<tr>
<td>3</td>
<td>0.999950</td>
<td>0.999950</td>
<td>0.999950</td>
<td>$4 \times 10^{-11}$</td>
<td>0.999950</td>
<td>0.999950</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\hat{P}(S \leq \tau)$</th>
<th>Glaz et al.</th>
<th>Our Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>0.712750</td>
<td>0.705787</td>
<td>0.714699</td>
<td>0.039308</td>
<td>0.697431</td>
<td>0.706948</td>
</tr>
<tr>
<td>14</td>
<td>0.867498</td>
<td>0.862184</td>
<td>0.865029</td>
<td>0.012502</td>
<td>0.859543</td>
<td>0.862407</td>
</tr>
<tr>
<td>15</td>
<td>0.946912</td>
<td>0.943329</td>
<td>0.946177</td>
<td>0.004169</td>
<td>0.942552</td>
<td>0.943362</td>
</tr>
<tr>
<td>16</td>
<td>0.980230</td>
<td>0.978959</td>
<td>0.979822</td>
<td>0.001354</td>
<td>0.978733</td>
<td>0.978963</td>
</tr>
<tr>
<td>17</td>
<td>0.993486</td>
<td>0.992821</td>
<td>0.993134</td>
<td>0.000433</td>
<td>0.992756</td>
<td>0.992822</td>
</tr>
<tr>
<td>18</td>
<td>0.997802</td>
<td>0.997726</td>
<td>0.997849</td>
<td>0.000127</td>
<td>0.997708</td>
<td>0.997726</td>
</tr>
<tr>
<td>19</td>
<td>0.999362</td>
<td>0.999327</td>
<td>0.999358</td>
<td>$3 \times 10^{-5}$</td>
<td>0.999322</td>
<td>0.999327</td>
</tr>
<tr>
<td>20</td>
<td>0.999819</td>
<td>0.999813</td>
<td>0.999825</td>
<td>$9 \times 10^{-6}$</td>
<td>0.999812</td>
<td>0.999813</td>
</tr>
<tr>
<td>21</td>
<td>0.999954</td>
<td>0.999951</td>
<td>0.999953</td>
<td>$2 \times 10^{-6}$</td>
<td>0.999951</td>
<td>0.999951</td>
</tr>
</tbody>
</table>
### One Dimensional Case: $X_i \sim B(n, p)$

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

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>Exact</th>
<th>Glaz et al. Product-type</th>
<th>Our Approximation</th>
<th>Approximation Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.810209</td>
<td>0.810216</td>
<td>0.810404</td>
<td>0.001111</td>
<td>0.809903</td>
<td>0.810439</td>
</tr>
<tr>
<td>2</td>
<td>0.995764</td>
<td>0.995764</td>
<td>0.995764</td>
<td>$3 \times 10^{-7}$</td>
<td>0.995764</td>
<td>0.995764</td>
</tr>
<tr>
<td>3</td>
<td>0.999950</td>
<td>0.999950</td>
<td>0.999950</td>
<td>$4 \times 10^{-11}$</td>
<td>0.999950</td>
<td>0.999950</td>
</tr>
</tbody>
</table>

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

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

**One dimensional case:** \( X_i \sim \mathcal{B}(n, p) \)

**Table 1:** \( n = 1, p = 0.005, m_1 = 10, T_1 = 1000, l t_{App} = 10^4 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>Exact</th>
<th>Glaz et al. Product-type</th>
<th>Our Approximation</th>
<th>Approximation Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.810209</td>
<td>0.810216</td>
<td>0.810404</td>
<td>0.001111</td>
<td>0.809903</td>
<td>0.810439</td>
</tr>
<tr>
<td>2</td>
<td>0.995764</td>
<td>0.995764</td>
<td>0.995764</td>
<td>3 \times 10^{-7}</td>
<td>0.995764</td>
<td>0.995764</td>
</tr>
<tr>
<td>3</td>
<td>0.999950</td>
<td>0.999950</td>
<td>0.999950</td>
<td>4 \times 10^{-11}</td>
<td>0.999950</td>
<td>0.999950</td>
</tr>
</tbody>
</table>

**Table 2:** \( n = 5, p = 0.05, m_1 = 25, T_1 = 500, l t_{App} = 10^4, l t_{Sim} = 10^3 \)

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

**Other numerical results**

A. Amărioarei (INRIA) 1d and 2d Scan Statistics SSB Seminar
Two dimensional case: $X_{i_1,i_2} \sim \mathcal{B}(n, p)$

**Table 3**: $n = 1, p = 0.005, m_1 = m_2 = 6, T_1 = T_2 = 30, \text{lt}_{App} = 10^3, \text{lt}_{Sim} = 10^3$

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\hat{P}(S \leq \tau)$</th>
<th>Glaz et al. Product-type Approximation</th>
<th>Our Product-type Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.915903</td>
<td>0.914013</td>
<td>0.920211</td>
<td>0.041483</td>
<td>0.901935</td>
<td>0.945623</td>
</tr>
<tr>
<td>3</td>
<td>0.994292</td>
<td>0.994395</td>
<td>0.994578</td>
<td>0.000803</td>
<td>0.993785</td>
<td>0.996638</td>
</tr>
<tr>
<td>4</td>
<td>0.999747</td>
<td>0.999757</td>
<td>0.999760</td>
<td>$2 \times 10^{-5}$</td>
<td>0.999737</td>
<td>0.999858</td>
</tr>
<tr>
<td>5</td>
<td>0.999992</td>
<td>0.999992</td>
<td>0.999992</td>
<td>$7 \times 10^{-7}$</td>
<td>0.999992</td>
<td>0.999995</td>
</tr>
</tbody>
</table>

**Table 4**: $n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, \text{lt}_{App} = 10^4$

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\hat{P}(S \leq \tau)$</th>
<th>Glaz et al. Product-type Approximation</th>
<th>Our Product-type Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.894654</td>
<td>0.873256</td>
<td>0.893724</td>
<td>0.037136</td>
<td>0.803422</td>
<td>0.944318</td>
</tr>
<tr>
<td>5</td>
<td>0.988003</td>
<td>0.986249</td>
<td>0.988144</td>
<td>0.002125</td>
<td>0.981418</td>
<td>0.993451</td>
</tr>
<tr>
<td>6</td>
<td>0.998963</td>
<td>0.998847</td>
<td>0.998963</td>
<td>0.000152</td>
<td>0.998543</td>
<td>0.999401</td>
</tr>
<tr>
<td>7</td>
<td>0.999926</td>
<td>0.999919</td>
<td>0.999925</td>
<td>$9 \times 10^{-6}$</td>
<td>0.999903</td>
<td>0.999955</td>
</tr>
<tr>
<td>8</td>
<td>0.999995</td>
<td>0.999995</td>
<td>0.999995</td>
<td>$5 \times 10^{-7}$</td>
<td>0.999994</td>
<td>0.999997</td>
</tr>
</tbody>
</table>
TWO DIMENSIONAL CASE: \( X_{i_1,i_2} \sim \mathcal{B}(n, p) \)

**Table 3:** \( n = 1, p = 0.005, m_1 = m_2 = 6, T_1 = T_2 = 30, l_{t_{App}} = 10^3, l_{t_{Sim}} = 10^3 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( \hat{P}(S \leq \tau) )</th>
<th>Glaz et al.</th>
<th>Our Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.915903</td>
<td>0.914013</td>
<td>0.920211</td>
<td>0.041483</td>
<td>0.901935</td>
<td>0.945623</td>
</tr>
<tr>
<td>3</td>
<td>0.994292</td>
<td>0.994395</td>
<td>0.994578</td>
<td>0.000803</td>
<td>0.993785</td>
<td>0.996638</td>
</tr>
<tr>
<td>4</td>
<td>0.999747</td>
<td>0.999757</td>
<td>0.999760</td>
<td>( 2 \times 10^{-5} )</td>
<td>0.999737</td>
<td>0.999858</td>
</tr>
<tr>
<td>5</td>
<td>0.999992</td>
<td>0.999992</td>
<td>0.999992</td>
<td>( 7 \times 10^{-7} )</td>
<td>0.999992</td>
<td>0.999995</td>
</tr>
</tbody>
</table>

**Table 4:** \( n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, l_{t_{App}} = 10^4 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( \hat{P}(S \leq \tau) )</th>
<th>Glaz et al.</th>
<th>Our Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.894654</td>
<td>0.873256</td>
<td>0.893724</td>
<td>0.037136</td>
<td>0.803422</td>
<td>0.944318</td>
</tr>
<tr>
<td>5</td>
<td>0.988003</td>
<td>0.986249</td>
<td>0.988144</td>
<td>0.002125</td>
<td>0.981418</td>
<td>0.993451</td>
</tr>
<tr>
<td>6</td>
<td>0.998963</td>
<td>0.998847</td>
<td>0.998963</td>
<td>0.000152</td>
<td>0.998543</td>
<td>0.999401</td>
</tr>
<tr>
<td>7</td>
<td>0.999926</td>
<td>0.999919</td>
<td>0.999925</td>
<td>( 9 \times 10^{-6} )</td>
<td>0.999903</td>
<td>0.999955</td>
</tr>
<tr>
<td>8</td>
<td>0.999995</td>
<td>0.999995</td>
<td>0.999995</td>
<td>( 5 \times 10^{-7} )</td>
<td>0.999994</td>
<td>0.999997</td>
</tr>
</tbody>
</table>
TWO DIMENSIONAL CASE: $X_{i_1,i_2} \sim \mathcal{B}(n, p)$

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

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\hat{P}(S \leq \tau)$</th>
<th>Glaz et al. Approximation</th>
<th>Our Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.915903</td>
<td>0.914013</td>
<td>0.920211</td>
<td>0.041483</td>
<td>0.901935</td>
<td>0.945623</td>
</tr>
<tr>
<td>3</td>
<td>0.994292</td>
<td>0.994395</td>
<td>0.994578</td>
<td>0.000803</td>
<td>0.993785</td>
<td>0.996638</td>
</tr>
<tr>
<td>4</td>
<td>0.999747</td>
<td>0.999757</td>
<td>0.999760</td>
<td>$2 \times 10^{-5}$</td>
<td>0.999737</td>
<td>0.999858</td>
</tr>
<tr>
<td>5</td>
<td>0.999992</td>
<td>0.999992</td>
<td>0.999992</td>
<td>$7 \times 10^{-7}$</td>
<td>0.999992</td>
<td>0.999995</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\hat{P}(S \leq \tau)$</th>
<th>Glaz et al. Approximation</th>
<th>Our Approximation</th>
<th>Total Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.894654</td>
<td>0.873256</td>
<td>0.893724</td>
<td>0.037136</td>
<td>0.803422</td>
<td>0.944318</td>
</tr>
<tr>
<td>5</td>
<td>0.988003</td>
<td>0.986249</td>
<td>0.988144</td>
<td>0.002125</td>
<td>0.981418</td>
<td>0.993451</td>
</tr>
<tr>
<td>6</td>
<td>0.998963</td>
<td>0.998847</td>
<td>0.998963</td>
<td>0.000152</td>
<td>0.998543</td>
<td>0.999401</td>
</tr>
<tr>
<td>7</td>
<td>0.999926</td>
<td>0.999919</td>
<td>0.999925</td>
<td>$9 \times 10^{-6}$</td>
<td>0.999903</td>
<td>0.999955</td>
</tr>
<tr>
<td>8</td>
<td>0.999995</td>
<td>0.999995</td>
<td>0.999995</td>
<td>$5 \times 10^{-7}$</td>
<td>0.999994</td>
<td>0.999997</td>
</tr>
</tbody>
</table>
MATLAB GUI application
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
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 \(f : S^k_Y \rightarrow S_W\) such that

\[
Z_n = f (Y_n, Y_{n+1}, \ldots, Y_{n+k-1}), \forall n \geq 1.
\]

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

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

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

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

Let $\Pi : \mathbb{R}^{c_1 \times c_2} \rightarrow \mathbb{R}$ 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

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

[Amărioarei and Preda, 2013b] and [Amărioarei and Preda, 2014]
**Examples for One and Two Dimensions**

**Example (One dimensional case)**

\[
X_{s_1} = \begin{bmatrix} \tilde{X}_{s_1-x_1(1)} & \cdots & \tilde{X}_{s_1-x_1(1)+x_2(1)} \end{bmatrix}
\]

\[
X_{s_1} = \prod \left( X_{s_1+x_1(1)} \right) = \prod \left( \tilde{X}_{s_1}, \ldots, \tilde{X}_{s_1+c_1-1} \right)
\]

**Example (Two dimensional case)**

\[
x_{s_1, s_2} = \begin{bmatrix} \tilde{X}_{s_1-x_1(1), s_2-x_1(2)} & \cdots & \tilde{X}_{s_1-x_1(1)+x_2(1), s_2-x_1(2)} \\ \vdots & \ddots & \vdots \\ \tilde{X}_{s_1-x_1(1), s_2+x_2(2)} & \cdots & \tilde{X}_{s_1-x_1(1)+x_2(1), s_2+x_2(2)} \end{bmatrix}
\]

\[
x_{s_1, s_2} = \prod \left( X_{s_1+x_1(1), s_2+x_2(1)} \right)
\]
Dependency structure in two dimensions

\[ \tilde{X}_{s_1-x_1^{(1)},s_2+x_2^{(2)}} \]

\[ \tilde{X}_{s_1+x_2^{(1)},s_2+x_2^{(2)}} \]

\[ \ldots \]

\[ \Pi \]

\[ X_{s_1-x_1^{(1)},s_2-x_1^{(2)}} \]
DEPENDENCY STRUCTURE IN TWO DIMENSIONS

\[ \tilde{X}_{s_1-x_1(1),s_2+x_2(2)} \]

\[ \tilde{X}_{s_1+x_2(1),s_2-x_2(2)} \]

\[ \tilde{T}_2 \]

\[ X_{s_1-x_1(1),s_2-x_2(2)} \]

\[ \Pi \]

\[ T_2 \]

\[ s_2 \]

\[ 1 \]

A. Amário arei (INRIA)
**Approximation: Idea**

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

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

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

- \((Z_{k_1})_{k_1}\) is 1-dependent, stationary and

\[
S_m(T) = \max_{1 \leq k_1 \leq L_1 - 1} Z_{k_1}
\]

**Illustration of the 1-dependence structure in two dimensions**

- Diagram showing the relationship between variables and their dependencies.
**Approximation process in two dimensions**

\[
T_2 \quad R_2
\]

\[
T_1
\]

\[
m_1
\]

\[
m_2
\]

\[
Q_{22} \quad 2m_1 + c_1 - 3
\]

\[
Q_{23} \quad 3m_2 + c_2 - 3
\]

\[
Q_{32} \quad 2m_2 + c_2 - 3
\]

\[
Q_{33} \quad 3m_2 + c_2 - 5
\]

\[
2m_1 + c_1 - 3
\]

\[
3m_1 + 2c_1 - 5
\]

**Error bounds**

A. Amărioarei (INRIA)  
1d and 2d Scan Statistics  
SSB Seminar 47 / 62
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 \( (\tilde{X}_n)_{n \geq 1} \) be a sequence of i.i.d. r.v.'s with the common distribution \( G \).

**Increasing Run**

A subsequence \( (\tilde{X}_k, \ldots, \tilde{X}_{k+l-1}) \) 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: $X_{s_1} = 1_{\tilde{X}_{s_1} < \tilde{X}_{s_1+1}}$

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

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

$X_{s_1} :$

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: \( X_{s_1} = 1_{\tilde{x}_{s_1} < \tilde{x}_{s_1+1}} \)

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

\[
\tilde{X}_{s_1} : \begin{array}{cccccccccc}
0.79 & 0.31 & 0.52 & 0.16 & 0.60 & 0.26 & 0.65 & 0.68 & 0.74 & 0.45
\end{array}
\]

\[
X_{s_1} : \begin{array}{c}
0
\end{array}
\]

We have

\[
\mathbb{P}(M_{\tilde{T}_1} \leq m_1) = \mathbb{P}(L_{T_1} < m_1) = \mathbb{P}(S_{m_1}(T_1) < m_1), \text{ for } m_1 \geq 1
\]
LONGEST INCREASING RUN

SCAN STATISTICS APPROACH

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

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

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

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

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

$X_{s_1} : \quad 0 \quad 1$

We have

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

Scan Statistics Approach

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

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

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

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

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

\(
X_{s_1} : \quad 0 \quad 1 \quad 0
\)

We have

\[
\mathbb{P}(M_{\tilde{T}_1} \leq m_1) = \mathbb{P}(L_{T_1} < m_1) = \mathbb{P}(S_{m_1}(T_1) < m_1), \text{ for } m_1 \geq 1
\]
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: $X_{s_1} = 1_{\tilde{x}_{s_1} < \tilde{x}_{s_1} + 1}$

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} : \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow \downarrow $

We have

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

SCAN STATISTICS APPROACH

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

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

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

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

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

$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

$$X_{s_1} : \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0$$

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

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

<table>
<thead>
<tr>
<th>$m_1$</th>
<th>Sim</th>
<th>AppH</th>
<th>$E_{total}(1)$</th>
<th>LimApp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.00000700</td>
<td>0.00000733</td>
<td>0.14860299</td>
<td>0.00000676</td>
</tr>
<tr>
<td>6</td>
<td>0.17567262</td>
<td>0.17937645</td>
<td>0.01089628</td>
<td>0.17620431</td>
</tr>
<tr>
<td>7</td>
<td>0.80257424</td>
<td>0.80362353</td>
<td>0.00110990</td>
<td>0.80215088</td>
</tr>
<tr>
<td>8</td>
<td>0.97548510</td>
<td>0.97566460</td>
<td>0.00011579</td>
<td>0.97550345</td>
</tr>
<tr>
<td>9</td>
<td>0.99749821</td>
<td>0.99751049</td>
<td>0.00000114</td>
<td>0.99749792</td>
</tr>
<tr>
<td>10</td>
<td>0.99977074</td>
<td>0.99977183</td>
<td>0.00000098</td>
<td>0.99977038</td>
</tr>
<tr>
<td>11</td>
<td>0.99998075</td>
<td>0.99998083</td>
<td>0.00000008</td>
<td>0.99998073</td>
</tr>
<tr>
<td>12</td>
<td>0.99999851</td>
<td>0.99999851</td>
<td>0.00000001</td>
<td>0.99999851</td>
</tr>
<tr>
<td>13</td>
<td>0.99999989</td>
<td>0.99999989</td>
<td>0.00000000</td>
<td>0.99999989</td>
</tr>
<tr>
<td>14</td>
<td>0.99999999</td>
<td>0.99999999</td>
<td>0.00000000</td>
<td>0.99999999</td>
</tr>
<tr>
<td>15</td>
<td>1.00000000</td>
<td>1.00000000</td>
<td>0.00000000</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>
Longest increasing run: numerical results

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

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

<table>
<thead>
<tr>
<th>$m_1$</th>
<th>Sim</th>
<th>AppH</th>
<th>$E_{total}(1)$</th>
<th>LimApp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.00000700</td>
<td>0.00000733</td>
<td>0.14860299</td>
<td>0.00000676</td>
</tr>
<tr>
<td>6</td>
<td>0.17567262</td>
<td>0.17937645</td>
<td>0.01089628</td>
<td>0.17620431</td>
</tr>
<tr>
<td>7</td>
<td>0.80257424</td>
<td>0.80362353</td>
<td>0.00110990</td>
<td>0.80215088</td>
</tr>
<tr>
<td>8</td>
<td>0.97548510</td>
<td>0.97566460</td>
<td>0.00011579</td>
<td>0.97550345</td>
</tr>
<tr>
<td>9</td>
<td>0.99749821</td>
<td>0.99751049</td>
<td>0.00001114</td>
<td>0.99749792</td>
</tr>
<tr>
<td>10</td>
<td>0.99977074</td>
<td>0.99977183</td>
<td>0.00000098</td>
<td>0.99977038</td>
</tr>
<tr>
<td>11</td>
<td>0.99998075</td>
<td>0.99998083</td>
<td>0.00000008</td>
<td>0.99998073</td>
</tr>
<tr>
<td>12</td>
<td>0.99999851</td>
<td>0.99999851</td>
<td>0.00000001</td>
<td>0.99999851</td>
</tr>
<tr>
<td>13</td>
<td>0.99999989</td>
<td>0.99999989</td>
<td>0.00000000</td>
<td>0.99999989</td>
</tr>
<tr>
<td>14</td>
<td>0.99999999</td>
<td>0.99999999</td>
<td>0.00000000</td>
<td>0.99999999</td>
</tr>
<tr>
<td>15</td>
<td>1.00000000</td>
<td>1.00000000</td>
<td>0.00000000</td>
<td>1.00000000</td>
</tr>
</tbody>
</table>
LONGEST INCREASING RUN: NUMERICAL RESULTS

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

\[
\mathbb{P}(M_{T_1} \leq m_1) \sim \exp(-\exp \eta),
\]

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

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

<table>
<thead>
<tr>
<th>( m_1 )</th>
<th>Sim</th>
<th>AppH</th>
<th>( E_{total}(1) )</th>
<th>LimApp</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.56445934</td>
<td>0.56997462</td>
<td>0.00255592</td>
<td>0.56810748</td>
</tr>
<tr>
<td>7</td>
<td>0.95295406</td>
<td>0.95325180</td>
<td>0.00018554</td>
<td>0.95294598</td>
</tr>
<tr>
<td>8</td>
<td>0.99658057</td>
<td>0.99659071</td>
<td>0.00001214</td>
<td>0.99657969</td>
</tr>
<tr>
<td>9</td>
<td>0.99979460</td>
<td>0.99979550</td>
<td>0.00000068</td>
<td>0.99979435</td>
</tr>
<tr>
<td>10</td>
<td>0.99998950</td>
<td>0.99998950</td>
<td>0.00000003</td>
<td>0.99998947</td>
</tr>
</tbody>
</table>

We used \( T_1 = 10000 \), \( p = 0.1 \) and \( \text{Iter} = 10^5 \).
Longest Increasing Run: Numerical Results

For $\tilde{X}_{s1} \sim \text{Geom}(p)$, [Louchard and Prodinger, 2003] showed that

$$\mathbb{P}(M_{T1} \leq m_1) \sim \exp(-\exp \eta),$$

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

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

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

We used $T_1 = 10000$, $p = 0.1$ and $Iter = 10^5$. 
**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
**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 \( \left( X_n \right)_{n \geq 1} \) is said to be an *moving average of order q* (MA(q)) if

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

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

Let \( (\tilde{X}_n)_{n \geq 1} \) 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

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

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 $s_1 \in \{1, \ldots, T_1\}$, the 1-way tensor $X_{s_1}$

$$X_{s_1} = (\tilde{X}_{s_1}, \tilde{X}_{s_1+1}, \ldots, \tilde{X}_{s_1+q})$$

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

$$\Pi(x_1, \ldots, x_{q+1}) = a_1 x_1 + a_2 x_2 + \cdots + a_{q+1} x_{q+1}.$$

Example (MA(2))

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

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

**Moving average of order $q$: numerical results**

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>Sim</th>
<th>AppPT</th>
<th>AppH</th>
<th>$E_{sapp}(1)$</th>
<th>$E_{sf}(1)$</th>
<th>$E_{total}(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.582252</td>
<td>0.589479</td>
<td>0.584355</td>
<td>0.011503</td>
<td>0.003653</td>
<td>0.015156</td>
</tr>
<tr>
<td>12</td>
<td>0.770971</td>
<td>0.773700</td>
<td>0.771446</td>
<td>0.002319</td>
<td>0.001691</td>
<td>0.004010</td>
</tr>
<tr>
<td>13</td>
<td>0.889986</td>
<td>0.890009</td>
<td>0.889431</td>
<td>0.000434</td>
<td>0.000733</td>
<td>0.001167</td>
</tr>
<tr>
<td>14</td>
<td>0.951529</td>
<td>0.954536</td>
<td>0.951723</td>
<td>0.000073</td>
<td>0.000297</td>
<td>0.000370</td>
</tr>
<tr>
<td>15</td>
<td>0.980653</td>
<td>0.982433</td>
<td>0.980675</td>
<td>0.000011</td>
<td>0.000113</td>
<td>0.000124</td>
</tr>
<tr>
<td>16</td>
<td>0.992827</td>
<td>0.993690</td>
<td>0.992791</td>
<td>0.000001</td>
<td>0.000040</td>
<td>0.000042</td>
</tr>
<tr>
<td>17</td>
<td>0.997486</td>
<td>0.995471</td>
<td>0.997499</td>
<td>0.000000</td>
<td>0.000013</td>
<td>0.000014</td>
</tr>
<tr>
<td>18</td>
<td>0.999186</td>
<td>0.999411</td>
<td>0.999188</td>
<td>0.000000</td>
<td>0.000004</td>
<td>0.000004</td>
</tr>
<tr>
<td>19</td>
<td>0.999754</td>
<td>0.999717</td>
<td>0.999754</td>
<td>0.000000</td>
<td>0.000001</td>
<td>0.000001</td>
</tr>
<tr>
<td>20</td>
<td>0.999930</td>
<td>1.000000</td>
<td>0.999930</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

![Graph showing the comparison between App and Sim results for various $n$ values.](image)
## Moving average of order $q$: Numerical Results

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

![Graph showing error bounds and numerical results vs $n$.]
Conclusions and Perspectives

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!


Reliability approximation for Markov chain imbeddable systems.

Bounds for the distribution of two-dimensional binary scan statistics.

Two-dimensional discrete scan statistics.

Approximations and inequalities for the distribution of a scan statistic for 0-1 Bernoulli trials.

Birkhäuser Boston, Inc., Boston.
On the length of the longest monotone block.
*Studio Scientiarum Mathematicarum Hungarica*, 31:35–46.

*Non uniform random variate generation.*
Springer-Verlag, New York.

Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas.

A note on runs of geometrically distributed random variables.


References

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

Distribution of the scan statistic for a sequence of bistate trials.

World Scientific Publishing Co., Inc., River Edge, NJ.

An efficient algorithm for exact distribution of discrete scan statistics.

Computation of Multivariate Normal and T Probabilities.
Springer-Verlag, New York.

A comparison of product-type and Bonferroni-type inequalities in presence of dependence.


*Extremes, 3:349–361.*


References

*Poisson processes.*
Oxford University Press.

Ascending runs of sequences of geometrically distributed random variables: a probabilistic analysis.

A compresive method for genome scans.

Computing scan statistic $p$ values using importance sampling, with applications to genetics and medical image analysis.

Abstract tubes, improved inclusion exclusion identities and inequalities and importance sampling.


Three problems on the length of increasing runs.

Importance sampling for estimating $p$ values in linkage analysis.

*Scan statistics for normal data.*

A variable window scan statistic for $MA(1)$ process.

Approximations and inequalities for moving sums.
On finite Markov chain imbedding technique.


A. Amãrioarei (INRIA)
PRODUCT-TYPE APPROXIMATION AND BOUNDS $d = 1$

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

- **Lower Bounds**
  \[ P(S_{m_1}(T_1) \leq \tau) \leq \frac{Q(2m_1)}{1 + \frac{Q(2m_1 - 1) - Q(2m_1)}{Q(2m_1) - 1} Q(2m_1)} T_1^{-2m_1}, \quad T_1 \geq 2m_1 \]
  \[ \leq \frac{Q(3m_1)}{1 + \frac{Q(2m_1 - 1) - Q(2m_1)}{Q(3m_1 - 1)} T_1^{-3m_1}}, \quad T_1 \geq 3m_1 \]

- **Upper Bounds**
  \[ P(S_{m_1}(T_1) \leq \tau) \leq Q(2m_1) [1 - Q(2m_1 - 1) + Q(2m_1)] T_1^{-2m_1}, \quad T_1 \geq 2m_1 \]
  \[ \leq Q(3m_1) [1 - Q(2m_1 - 1) + Q(2m_1)] T_1^{-3m_1}, \quad T_1 \geq 3m_1 \]

The values $Q(2m_1 - 1), Q(2m_1), Q(3m_1 - 1), Q(3m_1)$ are computed using [Karwe and Naus, 1997] algorithm.
PRODUCT-TYPE APPROXIMATION AND BOUNDS $d = 2$

- Approximation (Bernoulli)

\[ P(S_{m_1, m_2}(T_1, T_2) \leq k) \approx \frac{Q(m_1, m_2)(T_1 - m_1 - 1)(T_2 - m_2 - 1)Q(m_1 + 1, m_2 + 1)(T_1 - m_1)(T_2 - m_2)}{Q(m_1, m_2 + 1)(T_1 - m_1 - 1)(T_2 - m_2)Q(m_1 + 1, m_2)(T_1 - m_1)(T_2 - m_2 - 1)} \]

- Approximation (binomial and Poisson)

\[ P(S_{m_1, m_2}(T_1, T_2) \leq k) \approx \frac{Q(m_1 + 1, m_2 + 1)(T_1 - m_1)(T_2 - m_2)}{Q(m_1 + 1, m_2)(T_1 - m_1)(T_2 - m_2 - 1)} \times \frac{Q(m_1, 2m_2 - 1)(T_1 - m_1 - 1)(T_2 - 2m_2)}{Q(m_1, 2m_2)(T_1 - m_1 - 1)(T_2 - 2m_2 + 1)} \]

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
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

1. Define for $0 \leq k \leq m_1$
   $$F_{m_1,k} = \{ \Lambda_i | \Lambda_1 = 1 \ldots 1, \Lambda_2 = 1 0 1 \ldots 1, \ldots, \Lambda_l = 1 \ldots 1 0 \ldots 01 \}$$

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

2. The compound pattern $\Lambda = \bigcup_{i=1}^l \Lambda_i$, $\Lambda_i \in F_{m_1,k}$

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

   $$P(S_{m_1}(T_1) < k) = \xi N^{T_1} 1^\top, \text{where } \xi = (1, 0, \ldots, 0)$$
Appendix

Markov chain imbedding technique

Example

Consider the i.i.d. two-state sequence \((X_i)_{i \in \{1,2,\ldots,T_1\}}\) 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:
  
  \[
  N = \begin{pmatrix}
  0 & q & p & 0 & 0 & 0 & 0 \\
  0 & q & p & 0 & 0 & 0 & 0 \\
  0 & 0 & 0 & q & p & 0 & 0 \\
  0 & q & 0 & 0 & 0 & p & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & q \\
  0 & 0 & 0 & q & 0 & 0 & 0 \\
  0 & q & 0 & 0 & 0 & 0 & 0 
  \end{pmatrix}
  \]
### Selected Values for \( K(\cdot) \) and \( \Gamma(\cdot) \)

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

<table>
<thead>
<tr>
<th>( 1 - q_1 )</th>
<th>( K(1 - q_1) )</th>
<th>( \Gamma(1 - q_1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>38.63</td>
<td>480.69</td>
</tr>
<tr>
<td>0.05</td>
<td>21.28</td>
<td>180.53</td>
</tr>
<tr>
<td>0.025</td>
<td>17.56</td>
<td>145.20</td>
</tr>
<tr>
<td>0.01</td>
<td>15.92</td>
<td>131.43</td>
</tr>
</tbody>
</table>
Selected Values for $K(\cdot)$ and $\Gamma(\cdot)$

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

<table>
<thead>
<tr>
<th>$1 - q_1$</th>
<th>$K(1 - q_1)$</th>
<th>$\Gamma(1 - q_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>38.63</td>
<td>480.69</td>
</tr>
<tr>
<td>0.05</td>
<td>21.28</td>
<td>180.53</td>
</tr>
<tr>
<td>0.025</td>
<td>17.56</td>
<td>145.20</td>
</tr>
<tr>
<td>0.01</td>
<td>15.92</td>
<td>131.43</td>
</tr>
</tbody>
</table>
**Error bounds: approximation error**

**Approximation error**

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

where for \(2 \leq s \leq d\)

\[
F_{t_1, \ldots, t_{s-1}} = F \left( Q_{t_1, \ldots, t_{s-1}, 2}, L_s - 1 \right), \quad F = F (Q_2, L_1 - 1),
\]

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

\[
B_{t_1, \ldots, t_{d-1}} = (L_d - 1) F_{t_1, \ldots, t_{d-1}} \left( 1 - \gamma_{t_1, \ldots, t_{d-1}, 2} + B_{t_1, \ldots, t_{d-1}, 2} \right)^2, \quad B_{t_1, \ldots, t_d} = 0,
\]

and for \(s = 1\):

\[
\sum_{t_1, t_0 \in \{2,3\}} x = x, \quad F_{t_1, t_0} = F, \quad \gamma_{t_1, t_0, 2} = \gamma_2 \quad \text{and} \quad B_{t_1, t_0, 2} = B_2.
\]
ERROR BOUNDS: SIMULATION ERRORS

SIMULATION ERRORS

\[ E_{sf}(d) = (L_1 - 1) \cdots (L_d - 1) \sum_{t_1, \ldots, t_d \in \{2, 3\}} \beta_{t_1, \ldots, t_d} \]

\[ E_{sapp}(d) = \sum_{s=1}^{d} (L_1 - 1) \cdots (L_s - 1) \sum_{t_1, \ldots, t_{s-1} \in \{2, 3\}} F_{t_1, \ldots, t_{s-1}} \left( 1 - \hat{Q}_{t_1, \ldots, t_{s-1}, 2} \right) \]

\[ + A_{t_1, \ldots, t_{s-1}, 2} + C_{t_1, \ldots, t_{s-1}, 2} \]

\[ + \sum_{t_s \in \{2, 3\}} C_{t_1, \ldots, t_s} \]

where for \( 2 \leq s \leq d \)

\[ A_{t_1, \ldots, t_{s-1}} = (L_s - 1) \cdots (L_d - 1) \sum_{t_s, \ldots, t_d \in \{2, 3\}} \beta_{t_1, \ldots, t_d}, A_{t_1, \ldots, t_d} = \beta_{t_1, \ldots, t_d} \]

\[ C_{t_1, \ldots, t_{s-1}} = (L_s - 1) \left[ F_{t_1, \ldots, t_{s-1}} \left( 1 - \hat{Q}_{t_1, \ldots, t_{s-1}, 2} + A_{t_1, \ldots, t_{s-1}, 2} + C_{t_1, \ldots, t_{s-1}, 2} \right)^2 \right] \]

\[ + \sum_{t_s \in \{2, 3\}} C_{t_1, \ldots, t_s} \]
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 $\Sigma = (\Sigma_{i_1,j_1})$

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

**Step 2** requires to sample:

- $Y_{i_1^{(k)}}$ from the tail distribution $\mathbb{P}(Y_{i_1^{(k)}} \geq \tau)$ ([Devroye, 1986])

- for the other indices, from the conditional distribution given $\{Y_{i_1^{(k)}} \geq \tau\}$

For $W_1 = (Y_1, \ldots, Y_{i_1^{(k)}-1})$ and $W_2 = (Y_{i_1^{(k)}+1}, \ldots, Y_{T_1-m_1+1})$

\[
\overline{W}_1 = W_1 | (Y_{i_1^{(k)}} = t) \sim \mathcal{N}(\mu_{w_1 | t}, \Sigma_{w_1 | t}) \quad \text{and} \quad \overline{W}_2 = W_2 | (Y_{i_1^{(k)}} = t) \sim \mathcal{N}(\mu_{w_2 | t}, \Sigma_{w_2 | t})
\]

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

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

\[
\Sigma_{w_i | t} = \text{Cov}(W_i) - \frac{1}{\text{Var}[Y_{i_1^{(k)}}]} \text{Cov}[W_i, Y_{i_1^{(k)}}] \text{Cov}^T[W_i, Y_{i_1^{(k)}}].
\]
**Cumulative counts method**

**Idea**

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

- efficiently searches for the locality statistics over $R_d$ 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

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

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

\[
\text{Rel Eff} = \frac{\sigma^2_{\text{method 1}} \times \text{CPU Time}_{\text{method 1}}}{\sigma^2_{\text{method 2}} \times \text{CPU Time}_{\text{method 2}}}
\]
**Algorithm 3** Second Importance Sampling Algorithm for Scan Statistics

Take \( d_{\mathbb{P}}^\xi, r_1 = \frac{e^{\xi Y_{r_1}}}{E_{H_0} \left[ e^{\xi Y_{r_1}} \right]} d_{\mathbb{P}} H_0 \) and compute

\[
\xi \approx \frac{\mu}{m_1 \sigma^2} - \frac{\mu}{\sigma^2}, \quad E_{\xi, r_1} \left[ Y_{i_1} \right] = \xi \text{Cov}_{H_0} \left[ Y_{i_1}, Y_{r_1} \right] + m_1 \mu, \quad \text{Cov}_{\xi, r_1} \left[ Y_{i_1}, Y_{j_1} \right] = \text{Cov}_{H_0} \left[ Y_{i_1}, Y_{j_1} \right]
\]

Repeat for each \( k \) from 1 to \( \text{ITER} \) (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 \( \hat{\rho}_k(1) \) based on

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

End Repeat

Return

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

Comparison between methods

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)

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>$m_1$</th>
<th>$\tau$</th>
<th>Genz</th>
<th>Err Genz</th>
<th>IS Algo 2</th>
<th>Err Algo 2</th>
<th>Rel Eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>15</td>
<td>12</td>
<td>0.932483</td>
<td>0.000732</td>
<td>0.933215</td>
<td>0.000743</td>
<td>7</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>18</td>
<td>0.976117</td>
<td>0.000460</td>
<td>0.975797</td>
<td>0.000425</td>
<td>518</td>
</tr>
<tr>
<td>750</td>
<td>30</td>
<td>24</td>
<td>0.998454</td>
<td>0.000125</td>
<td>0.998493</td>
<td>0.000024</td>
<td>688</td>
</tr>
<tr>
<td>800</td>
<td>40</td>
<td>30</td>
<td>0.999752</td>
<td>0.000029</td>
<td>0.999742</td>
<td>0.000004</td>
<td>617</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>$m_1$</th>
<th>$\tau$</th>
<th>MC</th>
<th>Err MC</th>
<th>IS Algo 2</th>
<th>Err Algo 2</th>
<th>Rel Eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>15</td>
<td>12</td>
<td>0.932624</td>
<td>0.000694</td>
<td>0.933215</td>
<td>0.000743</td>
<td>15</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>18</td>
<td>0.975880</td>
<td>0.000425</td>
<td>0.975797</td>
<td>0.000425</td>
<td>33</td>
</tr>
<tr>
<td>750</td>
<td>30</td>
<td>24</td>
<td>0.998515</td>
<td>0.000061</td>
<td>0.998493</td>
<td>0.000024</td>
<td>101</td>
</tr>
<tr>
<td>800</td>
<td>40</td>
<td>30</td>
<td>0.999741</td>
<td>0.000009</td>
<td>0.999742</td>
<td>0.000004</td>
<td>602</td>
</tr>
</tbody>
</table>
### Numerical Results

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

<table>
<thead>
<tr>
<th>$T_1$</th>
<th>$m_1$</th>
<th>$\tau$</th>
<th>IS Algo 2</th>
<th>Err Algo 2</th>
<th>IS Algo 2</th>
<th>Err Algo 2</th>
<th>Rel Eff</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>15</td>
<td>12</td>
<td>0.932744</td>
<td>0.000839</td>
<td>0.933215</td>
<td>0.000743</td>
<td>3</td>
</tr>
<tr>
<td>500</td>
<td>25</td>
<td>18</td>
<td>0.976105</td>
<td>0.000448</td>
<td>0.975797</td>
<td>0.000425</td>
<td>3.5</td>
</tr>
<tr>
<td>750</td>
<td>30</td>
<td>24</td>
<td>0.998508</td>
<td>0.000032</td>
<td>0.998493</td>
<td>0.000024</td>
<td>3.5</td>
</tr>
<tr>
<td>800</td>
<td>40</td>
<td>30</td>
<td>0.999740</td>
<td>0.000006</td>
<td>0.999742</td>
<td>0.000004</td>
<td>3.6</td>
</tr>
</tbody>
</table>

**Figure 1**: The evolution of simulation error in IS Algorithm 2 and IS Algorithm 2
**Numerical results for big scanning window**

**Table 9**: \( n = 1, p = 0.01, m_1 = 10^4, T_1 = 10^6, l_{App} = 10^4 \)

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

**Table 9:** \( n = 1, p = 0.01, m_1 = 10^4, T_1 = 10^6, l_{App} = 10^4 \)

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

**Approximation error**

\[ E_{app}(d) = \sum_{s=1}^{d} (L_s - 1) \sum_{t_1, \ldots, t_{s-1} \in \{2,3\}} F_{t_1, \ldots, t_{s-1}} \left( 1 - \gamma_{t_1, \ldots, t_{s-1}, 2} + B_{t_1, \ldots, t_{s-1}, 2} \right)^2, \]

where for \( 2 \leq s \leq d \)

\[ F_{t_1, \ldots, t_{s-1}} = F\left(Q_{t_1, \ldots, t_{s-1}, 2}, L_s - 1\right), \]

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

\[ B_{t_1, \ldots, t_{d-1}} = (L_d - 1) F_{t_1, \ldots, t_{d-1}} \left( 1 - \gamma_{t_1, \ldots, t_{d-1}, 2} + B_{t_1, \ldots, t_{d-1}, 2} \right)^2, \]

and for \( s = 1 \):

\[ \sum_{t_1, t_0 \in \{2,3\}} x = x, \quad F_{t_1, t_0} = F, \quad \gamma_{t_1, t_0, 2} = \gamma_2 \text{ and } B_{t_1, t_0, 2} = B_2. \]
**ERROR BOUNDS: SIMULATION ERRORS**

**Simulation errors**

\[
E_{sf}(d) = (L_1 - 1) \cdot \ldots \cdot (L_d - 1) \sum_{t_1, \ldots, t_d \in \{2, 3\}} \beta_{t_1, \ldots, t_d}
\]

\[
E_{sapp}(d) = \sum_{s=1}^{d} (L_1 - 1) \cdot \ldots \cdot (L_s - 1) \sum_{t_1, \ldots, t_{s-1} \in \{2, 3\}} F_{t_1, \ldots, t_{s-1}} \left( 1 - \hat{Q}_{t_1, \ldots, t_{s-1}, 2} \right)
\]

\[
+ A_{t_1, \ldots, t_{s-1}, 2} + C_{t_1, \ldots, t_{s-1}, 2}^2
\]

where for \(2 \leq s \leq d\)

\[
A_{t_1, \ldots, t_{s-1}} = (L_s - 1) \cdot \ldots \cdot (L_d - 1) \sum_{t_s, \ldots, t_d \in \{2, 3\}} \beta_{t_1, \ldots, t_d}, \quad A_{t_1, \ldots, t_d} = \beta_{t_1, \ldots, t_d}
\]

\[
C_{t_1 \ldots t_{s-1}} = (L_s - 1) \left[ F_{t_1 \ldots t_{s-1}} \left( 1 - \hat{Q}_{t_1 \ldots t_{s-1}, 2} + A_{t_1 \ldots t_{s-1}, 2} + C_{t_1 \ldots t_{s-1}, 2} \right) \right]^2
\]

\[
+ \sum_{t_s \in \{2, 3\}} C_{t_1 \ldots t_s}
\]
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 \( \lambda \) and \( m_j \leq T_j, 1 \leq j \leq 2 \) be positive integers.

- Define for \( 0 \leq s_j \leq T_j - m_j \),

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

- The two dimensional scan statistic,

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

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

\[
\mathbb{P}(S_{m_1,m_2}(\lambda, T_1, T_2) \leq \tau) = \mathbb{P}\left(S_{1,1}(\lambda m_1 m_2, \frac{T_1}{m_1}, \frac{T_2}{m_2}) \leq \tau\right)
\]
Approximation process in two dimensions