Approximations for Multidimensional Discrete Scan Statistics

Doctoral Dissertation

Alexandru Amărioarei

Advisor: Cristian Preda

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

September 15, 2014
The one dimensional scan statistics

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 $X_{i_1} \sim B(p)$, $1 \leq i_1 \leq 20$)

| 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 \( m_1 \)
  \[
P(W_{m_1,k} \leq T_1) = P(S_{m_1}(T_1) \geq k)
  \]

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

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

**Problem**

Find a good estimate for the distribution of the discrete scan statistic \( P(S_m(T_1) \leq \tau) \).

**Previous work:**

- **One dimensional scan statistics**
  - Exact results ([Naus, 1974], [Fu, 2001], [Gao et al., 2005])
  - Approximations: product-type, Poisson ([Naus, 1982], [Chen and Glaz, 1997], [Glaz et al., 2001])
  - Bounds ([Glaz, 1990], [Glaz and Naus, 1991])

- **Two dimensional scan statistics**
  - Approximations: product-type, Poisson ([Chen and Glaz, 1996], [Boutsikas and Koutras, 2000])
  - Bounds ([Chen and Glaz, 1996], [Boutsikas and Koutras, 2003])

- **Three dimensional scan statistics**
  - Approximations: product-type, Poisson ([Guerriero et al., 2010])
The focus of this thesis

We consider the $d$ dimensional discrete scan statistics over a random field generated by:

- i.i.d. observations
- dependent (block-factor type) observations

We present

- accurate approximations
- error bounds
- simulation aspects
OUTLINE

1 Multidimensional Discrete Scan Statistics (i.i.d. model)
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2 Multidimensional Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Applications

3 Conclusions and Perspectives
   - Conclusions
   - Perspectives

4 References
OUTLINE

1. MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (I.I.D. MODEL)
   - Framework
     - Extremes of 1-dependent stationary sequences
     - Scan statistics and 1-dependent sequences
     - Simulation methods and computational aspects
     - Numerical examples

2. MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (BLOCK-FACTOR MODEL)
   - Model and discussion
   - Applications

3. CONCLUSIONS AND PERSPECTIVES
   - Conclusions
   - Perspectives

REFERENCES
Framework
**The $d$-dimensional discrete scan statistics**

Let $T_1$, $T_2$, $\ldots$, $T_d$ be positive integers, with $d \geq 1$

- The rectangular region, $\mathcal{R}_d = [0, T_1] \times [0, T_2] \times \cdots \times [0, T_d]$
- The r.v.’s $X_{s_1,s_2,\ldots,s_d}$, $1 \leq s_j \leq T_j$, $j \in \{1, 2, \ldots, d\}$

Let $2 \leq m_j \leq T_j$, $1 \leq j \leq d$, be positive integers
- Define for $1 \leq i_l \leq T_l - m_l + 1$, $1 \leq l \leq d$, $m = (m_1, m_2, \ldots, m_d)$ and $T = (T_1, T_2, \ldots, T_d)$

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

- The $d$-dimensional discrete scan statistic,

\[
S_m(T) = \max_{1 \leq i_j \leq T_j - m_j + 1, j \in \{1, 2, \ldots, d\}} Y_{i_1,i_2,\ldots,i_d}
\]
**Example: Two Dimensional Scan Statistics**

We have for $d = 2$

$$Y_{i_1,i_2} = \sum_{s_1 = i_1}^{i_1 + m_1 - 1} \sum_{s_2 = i_2}^{i_2 + m_2 - 1} X_{s_1,s_2}, \quad S_{m_1,m_2}(T_1, T_2) = \max_{1 \leq i_1 \leq T_1 - m_1 + 1, 1 \leq i_2 \leq T_2 - m_2 + 1} Y_{i_1,i_2}$$
Animation for 3 Dimensional Scan Statistics

Number of points: 0
OBJECTIVE

Find a good estimate for the distribution of $d$-dimensional discrete scan statistic

$$Q_m(T) = \mathbb{P}(S_m(T) \leq \tau)$$

The distribution of $S_m(T)$ is used for testing the null hypotheses of randomness against the alternative hypothesis of clustering.

EXAMPLE

Bernoulli model

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

$H_1$: There exists

$$\mathcal{R}(i_1, i_2, \ldots, i_d) = [i_1 - 1, i_1 + m_1 - 1] \times \cdots \times [i_d - 1, i_d + m_d - 1] \subset \mathcal{R}_d$$

where the r.v.'s $X_{s_1,s_2,...,s_d} \sim \mathcal{B}(p')$, $p' > p$ and $X_{s_1,s_2,...,s_d} \sim \mathcal{B}(p)$ outside $\mathcal{R}(i_1, i_2, \ldots, i_d)$
OUTLINE

1 MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (I.I.D. MODEL)
   - Framework
   - Extremes of 1-dependent stationary sequences
     - Scan statistics and 1-dependent sequences
     - Simulation methods and computational aspects
     - Numerical examples

2 MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (BLOCK-FACTOR MODEL)
   - Model and discussion
   - Applications

3 CONCLUSIONS AND PERSPECTIVES
   - Conclusions
   - Perspectives

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

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

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

Theorem [Amarioarei, 2012]

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

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

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

- 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

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

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

Theorem [Amárioarei, 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. **Multidimensional Discrete Scan Statistics (i.i.d. model)**
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2. **Multidimensional Discrete Scan Statistics (block-factor model)**
   - Model and discussion
   - Applications

3. **Conclusions and Perspectives**
   - Conclusions
   - 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]
$S_m(T)$ viewed as maximum of 1-dependent r.v.’s

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

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

  $Z_{k_1} = \max_{1 \leq i_1 \leq k_1(m_1 - 1)} \max_{1 \leq i_j \leq (L_j - 1)(m_j - 1)} Y_{i_1, i_2, \ldots, i_d}$

- $(Z_j)_j$ is 1-dependent and stationary

- Observe

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

**Example ($d = 1$)**

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

**Example** ($d = 3$)
Approximation process

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

$$Q_{t_1} = Q_{t_1}(\tau) = \mathbb{P}\left( \bigcap_{k_1=1}^{t_1-1} \{Z_{k_1} \leq \tau\} \right) = \mathbb{P}\left( \max_{1 \leq i_1 \leq (t_1-1)(m_1-1)} \max_{1 \leq i_j \leq (L_j-1)(m_j-1)} \sum_{j=1}^{d} Y_{i_1,i_2,\ldots,i_d} \leq \tau \right)$$

If $1 - Q_2 \leq 0.1$ then

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

Example ($d = 1$)

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

- Approximation error, about
  $$(L_1 - 1)F(Q_2, L_1 - 1)(1 - Q_2)^2$$
**Approximation process**

The approximation of $S_m(T)$ is an iterative process. The $s$ step, $1 \leq s \leq d$, becomes:

- Let

$$Q_{t_1, t_2, \ldots, t_s} = Q_{t_1, t_2, \ldots, t_s}(\tau) = \mathbb{P} \left( \max_{1 \leq i_l \leq (t_l - 1)(m_l - 1)} Y_{i_1, i_2, \ldots, i_d} \leq \tau \right)$$

- Define for $t_l \in \{2, 3\}$, $l \in \{1, \ldots, s - 1\}$ and $k_s \in \{1, 2, \ldots, L_s - 1\}$

$$Z_{k_s}^{(t_1, t_2, \ldots, t_{s-1})} = \max_{1 \leq i_l \leq (t_l - 1)(m_l - 1)} Y_{i_1, i_2, \ldots, i_d}$$

$$Z_{k_s} = \max_{1 \leq i_l \leq (t_l - 1)(m_l - 1)} Y_{i_1, i_2, \ldots, i_d}$$

$$\left\{ Z_1^{(t_1, t_2, \ldots, t_{s-1})}, \ldots, Z_{L_s-1}^{(t_1, t_2, \ldots, t_{s-1})} \right\} \text{ forms a 1-dependent stationary sequence}$$

- If we take $H(x, y, m) = \frac{2x-y}{[1+x-y+2(x-y)^2]m-1}$, then we have the approximation

$$|Q_{t_1, \ldots, t_{s-1}} - H\left(Q_{t_1, \ldots, t_{s-1}, 2}, Q_{t_1, \ldots, t_{s-1}, 3}, L_s\right)| \leq (L_s - 1) F(Q_{t_1, \ldots, t_{s-1}, 2}, L_s) (1 - Q_{t_1, \ldots, t_{s-1}, 2})^2$$
ILLUSTRATION FOR $d = 2$
ILLUSTRATION FOR $d = 3$
ILLUSTRATION FOR $d = 3$
Illustration for $d = 3$
ERROR BOUNDS

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

$$\gamma_{t_1}, \ldots, t_{s-1} = H(\gamma_{t_1}, \ldots, t_{s-1}, 2, \gamma_{t_1}, \ldots, t_{s-1}, 3, L_s), \ 2 \leq s \leq d$$

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

$$\hat{Q}_{t_1}, \ldots, t_{s-1} = H(\hat{Q}_{t_1}, \ldots, t_{s-1}, 2, \hat{Q}_{t_1}, \ldots, t_{s-1}, 3, L_s), \ 2 \leq s \leq d$$

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 $\hat{Q}_{t_1}, \ldots, t_d$ will be estimated by Monte Carlo simulations.
ERROR BOUNDS

Let \( \gamma_{t_1, \ldots, t_d} = Q_{t_1, \ldots, t_d} \), with \( t_j \in \{2, 3\}, j \in \{1, \ldots, d\} \), and define

\[
\gamma_{t_1, \ldots, t_{s-1}} = H(\gamma_{t_1, \ldots, t_{s-1}, 2}, \gamma_{t_1, \ldots, t_{s-1}, 3}, L_s), \quad 2 \leq s \leq d
\]

Denote with \( \hat{Q}_{t_1, \ldots, t_d} \) the estimated value of \( Q_{t_1, \ldots, t_d} \) and define

\[
\hat{Q}_{t_1, \ldots, t_{s-1}} = H(\hat{Q}_{t_1, \ldots, t_{s-1}, 2}, \hat{Q}_{t_1, \ldots, t_{s-1}, 3}, L_s), \quad 2 \leq s \leq d
\]

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 \( \hat{Q}_{t_1, \ldots, t_d} \) will be estimated by Monte Carlo simulations.
ERROR BOUNDS

Let \( \gamma_{t_1}, \ldots, t_d = Q_{t_1}, \ldots, t_d \), with \( t_j \in \{2, 3\}, j \in \{1, \ldots, d\} \), and define

\[
\gamma_{t_1}, \ldots, t_{s-1} = H(\gamma_{t_1}, \ldots, t_{s-1}, 2, \gamma_{t_1}, \ldots, t_{s-1}, 3, L_s), 2 \leq s \leq d
\]

Denote with \( \hat{Q}_{t_1}, \ldots, t_d \) the estimated value of \( Q_{t_1}, \ldots, t_d \) and define

\[
\hat{Q}_{t_1}, \ldots, t_{s-1} = H(\hat{Q}_{t_1}, \ldots, t_{s-1}, 2, \hat{Q}_{t_1}, \ldots, t_{s-1}, 3, L_s), 2 \leq s \leq d
\]

OBJECTIVE

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

We observe that

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

\[
E_{app}(d) \leq E_{sapp}(d) + E_{sf}(d)
\]

The quantities \( \hat{Q}_{t_1}, \ldots, t_d \) will be estimated by Monte Carlo simulations.
Outline

1. Multidimensional Discrete Scan Statistics (i.i.d. model)
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2. Multidimensional Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Applications

3. Conclusions and Perspectives
   - Conclusions
   - Perspectives

References
Simulation methods and computational aspects
**Naive Hit-or-Miss MC**

**Objective**

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

### Algorithm 1: Classical Monte Carlo algorithm for scan statistics

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

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

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 (especially for $d \geq 2$)
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

$$P_{H_0} (S_m(T) \geq \tau) = \mathbb{P} \left( \bigcup_{i_1=1}^{T_1-m_1+1} \bigcup_{i_d=1}^{T_d-m_d+1} E_{i_1,..,i_d} \right) = \int G(x) f(x) \, dx$$

where $$E_{i_1,..,i_d} = \left\{ Y_{i_1,..,i_d} \geq \tau \right\}$$, $$G(x) = 1_E(x)$$, $$E = \bigcup_{i_1=1}^{T_1-m_1+1} \bigcup_{i_d=1}^{T_d-m_d+1} E_{i_1,..,i_d}$$ and $$f$$ is the joint density of $$Y_{i_1,..,i_d}$$ under $$H_0$$. 
IMPORTANTANCE SAMPLING FOR SCAN STATISTICS

We introduce the change of measure

\[ g(x) = \sum_{j_1=1}^{T_1-m_1+1} \ldots \sum_{j_d=1}^{T_d-m_d+1} \left\{ \frac{\mathbb{P}(E_{j_1}, \ldots, j_d)}{B(d)} \right\} \left\{ \frac{1_{E_{j_1}, \ldots, j_d} f(x)}{\mathbb{P}(E_{j_1}, \ldots, j_d)} \right\} \]

and we observe that

\[ \mathbb{P}_H(S_m(T) \geq \tau) = B(d) \rho(d) \]

- the Bonferroni upper bound \( B(d) \)

\[ B(d) = \sum_{i_1=1}^{T_1-m_1+1} \ldots \sum_{i_d=1}^{T_d-m_d+1} \mathbb{P}(E_{i_1}, \ldots, j_d) \]

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

\[ \rho(d) = \sum_{i_1=1}^{T_1-m_1+1} \ldots \sum_{i_d=1}^{T_d-m_d+1} p_{i_1, \ldots, i_d} \int \frac{1}{C(Y)} d\mathbb{P}_H(\cdot \mid E_{i_1}, \ldots, i_d) \]

where

\[ p_{i_1, \ldots, i_d} = \frac{1}{(T_1-m_1+1) \ldots (T_d-m_d+1)} \]

\[ C(Y) = \sum_{i_1=1}^{T_1-m_1+1} \ldots \sum_{i_d=1}^{T_d-m_d+1} 1_{E_{i_1}, \ldots, i_d} \]
**Algorithm 2** Importance Sampling Algorithm for Scan Statistics

**Begin**

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

1: Generate uniformly the $d$-tuple $(i_1^{(k)}, \ldots, i_d^{(k)})$ from the set $\{1, \ldots, T_1 - m_1 + 1\} \times \cdots \times \{1, \ldots, T_d - m_d + 1\}$.

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

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

**End Repeat**

**Return**

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

**End**
IMPLEMENTATION PROBLEMS

Algorithm 2 presents two main difficulties:

A) being able to sample from the conditional distribution of \( X \) given
\[
\left\{ Y_{i_1^{(k)}}, \ldots, i_d^{(k)} \geq \tau \right\}
\]
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. MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (I.I.D. MODEL)
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2. MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (BLOCK-FACTOR MODEL)
   - Model and discussion
   - Applications

3. CONCLUSIONS AND PERSPECTIVES
   - Conclusions
   - Perspectives

REFERENCES
Numerical examples
# Examples for \( d = 1, 2, 3 \) When \( X_{i_1, \ldots, i_d} \sim \mathcal{B}(n, p) \)

## Table 1: \( n = 1, p = 0.005, m_1 = 10, T_1 = 1000, \text{lt}_{\text{App}} = 10^4 \)

<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: \( n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, \text{lt}_{\text{App}} = 10^4 \)

<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>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>

## Table 3: \( n = 1, p = 0.0001, m_1 = m_2 = m_3 = 5, T_1 = T_2 = T_3 = 60, \text{lt}_{\text{App}} = 10^5 \)

<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>Approximation Error</th>
<th>Simulation Error</th>
<th>Total Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.993294</td>
<td>0.993241</td>
<td>0.993192</td>
<td>0.000010</td>
<td>0.001367</td>
<td>0.001377</td>
</tr>
<tr>
<td>3</td>
<td>0.999963</td>
<td>0.999964</td>
<td>0.999963</td>
<td>0.000000</td>
<td>0.000005</td>
<td>0.000005</td>
</tr>
<tr>
<td>4</td>
<td>0.999999</td>
<td>0.999999</td>
<td>0.999999</td>
<td>0.000000</td>
<td>(2 \times 10^{-9})</td>
<td>(2 \times 10^{-9})</td>
</tr>
</tbody>
</table>
### Examples for $d = 1, 2, 3$ When $X_{i_1,...,i_d} \sim B(n, p)$

**Table 1:** $n = 1, p = 0.005, m_1 = 10, T_1 = 1000, \text{lt}_{App} = 10^4$

<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><strong>0.810404</strong></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><strong>0.999950</strong></td>
<td>$4 \times 10^{-11}$</td>
<td>0.999950</td>
<td>0.999950</td>
</tr>
</tbody>
</table>

**Table 2:** $n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, \text{lt}_{App} = 10^4$

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\hat{P}(S \leq \tau)$</th>
<th>Glaz et al. Product-type</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>

**Table 3:** $n = 1, p = 0.0001, m_1 = m_2 = m_3 = 5, T_1 = T_2 = T_3 = 60, \text{lt}_{App} = 10^5$

<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>Approximation Error</th>
<th>Simulation Error</th>
<th>Total Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.993294</td>
<td>0.993241</td>
<td>0.993192</td>
<td>0.000010</td>
<td>0.001367</td>
<td>0.001377</td>
</tr>
<tr>
<td>3</td>
<td>0.999963</td>
<td>0.999964</td>
<td>0.999963</td>
<td>0.000000</td>
<td>0.000005</td>
<td>0.000005</td>
</tr>
<tr>
<td>4</td>
<td>0.999999</td>
<td>0.999999</td>
<td>0.999999</td>
<td>0.000000</td>
<td>$2 \times 10^{-9}$</td>
<td>$2 \times 10^{-9}$</td>
</tr>
</tbody>
</table>
**Examples for** \( d = 1, 2, 3 \) **when** \( X_{i_1, \ldots, i_d} \sim \mathcal{B}(n, p) \)

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

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( \hat{P}(S \leq \tau) )</th>
<th>Glaz et al. Product-type</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>

**Table 3:** \( n = 1, p = 0.0001, m_1 = m_2 = m_3 = 5, T_1 = T_2 = T_3 = 60, l_{t_{App}} = 10^5 \)

<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>Approximation Error</th>
<th>Simulation Error</th>
<th>Total Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.993294</td>
<td>0.993241</td>
<td>0.993192</td>
<td>0.000010</td>
<td>0.001367</td>
<td>0.001377</td>
</tr>
<tr>
<td>3</td>
<td>0.999963</td>
<td>0.999964</td>
<td>0.999963</td>
<td>0.000000</td>
<td>0.000005</td>
<td>0.000005</td>
</tr>
<tr>
<td>4</td>
<td>0.999999</td>
<td>0.999999</td>
<td>0.999999</td>
<td>0.000000</td>
<td>( 2 \times 10^{-9} )</td>
<td>( 2 \times 10^{-9} )</td>
</tr>
</tbody>
</table>
EXAMPLES FOR \(d = 1, 2, 3\) WHEN \(X_{i_1,...,i_d} \sim \mathcal{B}(n, p)\)

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

| \(\tau\) | 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       | \(3 \times 10^{-7}\) | 0.995764         | 0.995764         |
| 3       | 0.999950                       | 0.999950        | 0.999950       | \(4 \times 10^{-11}\) | 0.999950         | 0.999950         |

**Table 2:** \(n = 5, p = 0.002, m_1 = 5, m_2 = 10, T_1 = 50, T_2 = 80, \text{lt}_{App} = 10^4\)

| \(\tau\) | \(\hat{P}(S \leq \tau)\) 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       | \(9 \times 10^{-6}\) | 0.999903         | 0.999955         |
| 8       | 0.999995                       | 0.999995        | 0.999995       | \(5 \times 10^{-7}\) | 0.999994         | 0.999997         |

**Table 3:** \(n = 1, p = 0.0001, m_1 = m_2 = m_3 = 5, T_1 = T_2 = T_3 = 60, \text{lt}_{App} = 10^5\)

| \(\tau\) | \(\hat{P}(S \leq \tau)\) Glaz et al. Product-type Our Approximation Approximation Error Simulation Error Total Error |
|---------|--------------------------------|-----------------|----------------|------------------|------------------|------------------|
| 2       | 0.993294                       | 0.993241        | 0.993192       | 0.000010         | 0.001367         | 0.001537         |
| 3       | 0.999963                       | 0.999964        | 0.999963       | 0.000000         | 0.000005         | 0.000005         |
| 4       | 0.999999                       | 0.999999        | 0.999999       | 0.000000         | \(2 \times 10^{-9}\) | \(2 \times 10^{-9}\) |
MATLAB GUI application
## OUTLINE

1. **Multidimensional Discrete Scan Statistics (i.i.d. model)**
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2. **Multidimensional Discrete Scan Statistics (block-factor model)**
   - Model and discussion
   - Applications

3. **Conclusions and Perspectives**
   - Conclusions
   - 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 $f : S_Y^k \rightarrow S_W$ 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 $1 \leq j \leq d$, $d \geq 1$, 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 $2 \leq m_j \leq T_j$, $1 \leq j \leq d$ be nonnegative integers.

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

To each $d$-tuple $(s_1, \ldots, s_d)$, with $s_j \in \{x_1^{(j)} + 1, \ldots, \tilde{T}_j - x_2^{(j)}\}$, $j \in \{1, \ldots, d\}$, associate a $d$-way tensor $X_{s_1, \ldots, s_d} \in \mathbb{R}^{c_1 \times \cdots \times c_d}$

\[
X_{s_1, \ldots, s_d}(j_1, \ldots, j_d) = \tilde{X}_{s_1 - x_1^{(1)} - 1 + j_1, \ldots, s_d - x_1^{(d)} - 1 + j_d}
\]

where $(j_1, \ldots, j_d) \in \{1, \ldots, c_1\} \times \cdots \times \{1, \ldots, c_d\}$.

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

\[
X_{s_1, \ldots, s_d} = \Pi \left( X_{s_1 + x_1^{(1)}, \ldots, s_d + x_1^{(d)}} \right)
\]

- for $d = 2$: [Amărioarei and Preda, 2013b] and [Amărioarei and Preda, 2014]
**Examples for One and Two Dimensions**

**Example \((d = 1)\)**

\[
X_{s_1} = \begin{bmatrix}
\tilde{X}_{s_1 - x_1^{(1)}}, \ldots, \tilde{X}_{s_1 + x_1^{(1)}}
\end{bmatrix}
\]

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

**Example \((d = 2)\)**

\[
X_{s_1, s_2} =
\begin{pmatrix}
\tilde{X}_{s_1 - x_1^{(1)}, s_2 - x_1^{(2)}} & \ldots & \tilde{X}_{s_1 + x_1^{(1)}, s_2 - x_1^{(2)}} \\
\vdots & \ddots & \vdots \\
\tilde{X}_{s_1 - x_1^{(1)}, s_2 + x_2^{(2)}} & \ldots & \tilde{X}_{s_1 + x_1^{(1)}, s_2 + x_2^{(2)}}
\end{pmatrix}
\]

\[
X_{s_1, s_2} = \prod \left( X_{s_1 + x_1^{(1)}, s_2 + x_2^{(2)}} \right)
\]
**Dependency structure \((d = 2)\)**
Dependency structure ($d = 2$)
**Approximation: Idea**

Let $L_j = \frac{\bar{\tau}_j}{m_j + c_j - 2}$, $j \in \{1, 2, \ldots, d\}$, be positive integers.

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

  $$Z_{k_1} = \max_{(k_1 - 1)(m_1 + c_1 - 2)+1 \leq i_1 \leq k_1 (m_1 + c_1 - 2)} \left\{ \max_{1 \leq i_j \leq (L_j - 1)(m_j + c_j - 2)} Y_{i_1, i_2, \ldots, i_d} \right\}$$

- $(Z_j)_j$ is 1-dependent, stationary and

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

**Example (1-dependence of $(Z_j)_j$ for $d = 2$)**

![Diagram showing the approximation idea for $d = 2$.]
Approximation process \((d = 2)\)
OUTLINE

1 Multidimensional Discrete Scan Statistics (i.i.d. model)
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2 Multidimensional Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Applications

3 Conclusions and Perspectives
   - Conclusions
   - Perspectives

REFERENCES
Applications
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

Let \( d = 1, \ c_1 = 2, \ T_1 = \tilde{T}_1 - 1 \) 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 (\( \tilde{X}_{s_1} \sim \mathcal{U}(0, 1), \ \tilde{T}_1 = 10 \))

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

\[ X_{s_1} : \]

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

Let $d = 1$, $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ 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: $X_{s_1} = 1_{X_{s_1} < X_{s_1+1}}$

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

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

$X_{s_1} : 0$

We have

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

SCAN STATISTICS APPROACH

Let \( d = 1, \ c_1 = 2, \ T_1 = \tilde{T}_1 - 1 \) 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 (\( \tilde{X}_{s_1} \sim \mathcal{U}(0, 1), \ \tilde{T}_1 = 10 \))

\[
\begin{align*}
\tilde{X}_{s_1} : & \quad 0.79 \quad 0.31 \quad 0.52 \quad 0.16 \quad 0.60 \quad 0.26 \quad 0.65 \quad 0.68 \quad 0.74 \quad 0.45 \\
X_{s_1} : & \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘} \quad \text{↘}
\end{align*}
\]

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

Let \( d = 1, c_1 = 2, T_1 = \tilde{T}_1 - 1 \) 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 (\( \tilde{X}_{s_1} \sim \mathcal{U}(0, 1), \tilde{T}_1 = 10 \))

\[\begin{array}{cccccccccc}
\tilde{X}_{s_1} & : & 0.79 & 0.31 & 0.52 & 0.16 & 0.60 & 0.26 & 0.65 & 0.68 & 0.74 & 0.45 \\
X_{s_1} & : & 0 & 1 & & & & & & & \\
\end{array}\]

We have

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

SCAN STATISTICS APPROACH

Let \( d = 1, \ c_1 = 2, \ T_1 = \tilde{T}_1 - 1 \) 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 (\( \tilde{X}_{s_1} \sim \mathcal{U}(0, 1), \ \tilde{T}_1 = 10 \))

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

\[
\begin{array}{cccccccccc}
X_{s_1} & : & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\
\end{array}
\]

We have

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

SCAN STATISTICS APPROACH

Let $d = 1$, $c_1 = 2$, $T_1 = \tilde{T}_1 - 1$ 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 ($\tilde{X}_{s_1} \sim \mathcal{U}(0, 1)$, $\tilde{T}_1 = 10$)

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

We have

$$\mathbb{P}
\left( M_{\tilde{T}_1} \leq m_1 \right) = \mathbb{P}
\left( L_{T_1} < m_1 \right) = \mathbb{P}
\left( S_{m_1}(T_1) < m_1 \right), \text{ for } m_1 \geq 1$$
For $\tilde{X}_{51} \sim \mathcal{U}([0, 1])$, [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>$E_{\text{total}}(1)$</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>0.97664640</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 $\tilde{X}_{s_1} \sim U([0,1])$, [Novak, 1992] showed that

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

<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>
**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 $q$**

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 $q$

**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 $\Pi : \mathbb{R}^{q+1} \to \mathbb{R}$

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

**Example (**MA$(2)$**))

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

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

- Product-type approximation for $MA(2)$: [Wang and Glaz, 2013] and [Wang, 2013].
## 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 $P(S_{m1}(T_1) \leq n)$ with error bounds for varying $n$.](image-url)
### 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>
OUTLINE

1 MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (I.I.D. MODEL)
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2 MULTIDIMENSIONAL DISCRETE SCAN STATISTICS (BLOCK-FACTOR MODEL)
   - Model and discussion
   - Applications

3 CONCLUSIONS AND PERSPECTIVES
   - Conclusions
   - Perspectives

REFERENCES
In this talk:

- improved a result concerning extremes of 1-dependent sequences
- introduced the multidimensional discrete scan statistics
- introduced a new model of dependence based on block-factor constructions
- presented a unified method for estimating the distribution of the multidimensional discrete scan statistics both for the i.i.d model and the block-factor model
- illustrated an importance sampling algorithm that increases the efficiency of the proposed approximation
OUTLINE

1 Multidimensional Discrete Scan Statistics (i.i.d. model)
   - Framework
   - Extremes of 1-dependent stationary sequences
   - Scan statistics and 1-dependent sequences
   - Simulation methods and computational aspects
   - Numerical examples

2 Multidimensional Discrete Scan Statistics (block-factor model)
   - Model and discussion
   - Applications

3 Conclusions and Perspectives
   - Conclusions
   - Perspectives

REFERENCES
**Future work**

Extend the results to

- multidimensional continuous scan statistics
- multidimensional conditional scan statistics

Investigate

- other dependent models
- the influence of the shape of the scanning window
- power of scan statistic based tests under different models
- scan statistics on graphs
*arXiv:1211.5456v1, submitted.*


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.

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

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

A note on runs of geometrically distributed random variables.


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.

*J. Appl. Probab., 38:908–916.*


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.

Tight bounds and approximations for scan statistic probabilities for discrete data.

*Scan statistics.*

Combinatorics of geometrically distributed random variables: run statistics.

Erdos-Révész type bounds for the length of the longest run from a stationary mixing sequence.
Approximations for a three dimensional scan statistic. 

First passage time for some stationary processes. 

Estimating the distributions of scan statistics with high precision. 

Estimating the distribution of one-dimensional discrete scan statistics viewed as extremes of 1-dependent stationary sequences. 

A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. 


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


*Scan statistics for normal data.*


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

thank you!
PRODUCT-TYPE APPROXIMATIONS

- One dimensional scan statistics

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

- Two dimensional scan statistics

\[ P(S_{m_1,m_2}(T_1, T_2) \leq \tau) \approx \frac{Q(m_1 + 1, m_2 + 1, m_3 + 1)}{Q(m_1 + 1, m_2)(T_1 - m_1)(T_2 - m_2)} \frac{Q(m_1 + 1, m_2)}{Q(m_1, 2m_2)(T_1 - m_1)(T_2 - m_2 - 1)} \times \frac{Q(m_1, 2m_2 - 1, m_3 + 1)}{Q(m_1, 2m_2)(T_1 - m_1 - 1)(T_2 - 2m_2)} \]

- Three dimensional scan statistics

\[ P(S_{m_1,m_2,m_3}(T_1, T_2, T_3) \leq \tau) \approx \frac{Q(m_1 + 1, m_2 + 1, m_3 + 1, m_4)}{Q(m_1, m_2, m_3)(T_1 - m_1)(T_2 - m_2)(T_3 - m_3)} \frac{Q(m_1 + 1, m_2 + 1, m_3 + 1)}{Q(m_1 + 1, m_2)(T_1 - m_1)(T_2 - 2m_2)} \times \frac{Q(m_1 + 1, m_2 + 1, m_3 + 1)}{Q(m_1, 2m_2)(T_1 - m_1 - 1)(T_2 - 2m_2 - 1)} \times \frac{Q(m_1, 2m_2 - 1, m_3 + 1)}{Q(m_1, 2m_2)(T_1 - m_1 - 1)(T_2 - 2m_2)} \]
### 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 4**: 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

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

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

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

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

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**

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

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

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

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

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

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

\[+ \sum_{t_s \in \{2, 3\}} C_{t_1, \ldots, t_s}\]
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 $\mu = m_1 \mu$ 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 $W_2 = (Y_{i_1(k)+1}, \ldots, Y_{T_1-m_1+1})$

$$W_1 = W_1|Y_{i_1(k)} = t) \sim \mathcal{N}(\mu_{W_1|t}, \Sigma_{W_1|t}) \text{ and } W_2 = W_2|Y_{i_1(k)} = t) \sim \mathcal{N}(\mu_{W_2|t}, \Sigma_{W_2|t})$$

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

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

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

Return
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 $\mathcal{R}_d$ 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

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

**Example: normal data**

**Alternative approaches**

Several other methods were proposed:

1) [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])

2) [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]

\[
\text{Rel Eff} = \frac{\sigma^2_{\text{method } 1} \times \text{CPU Time}_{\text{method } 1}}{\sigma^2_{\text{method } 2} \times \text{CPU Time}_{\text{method } 2}}
\]
IS algorithm [Shi et al., 2007]

Algorithm 3 Second Importance Sampling Algorithm for Scan Statistics

Take $dP_{\xi, r_1} = \frac{e^{\xi Y_{r_1}}}{E_{H_0}[e^{\xi Y_{r_1}}]} dP_{H_0}$ and compute

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

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 $dP_{\xi, i_1^{(k)}}$.

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

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

End Repeat

Return

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

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

**Table 5**: 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 6**: 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 7:** 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.
**Appendix**

**Approximation process block-factor model**

### 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 \(2 \leq s \leq d\)

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

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

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

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

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

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

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

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

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

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

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