Scan Statistics for Normal Data with Applications

Joseph Glaz

University of Connecticut

October 2013
Scan statistics for normal data.

- Introduction
- Probability Inequalities: one dimensional data.
- Product-type approximations for one dimensional data.
- Inequalities and approximations: two dimensional data.
- Variable window scan statistics: minimum p-value approach.
- Applications to time series data.

Summary and future work.
Introduction: Early References on Clustering of Events

Introduction: Early Theoretical Advances on Scan Statistics


Introduction: Early Theoretical Advances


Scan statistics are used for detecting clusters of rare events.
Scan statistics are used for detecting clusters of rare events.

Applications:
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
Scan statistics are used for detecting clusters of rare events.

Applications:
- Agricultural Sciences
- Astronomy
Introduction: Applications

- Scan statistics are used for detecting clusters of rare events.
- Applications:
  - Agricultural Sciences
  - Astronomy
  - Bioinformatics
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
Scan statistics are used for detecting clusters of rare events.

Applications:
- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
- Epidemiology
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
- Epidemiology
- Physics
Scan statistics are used for detecting clusters of rare events.

Applications:
- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
- Epidemiology
- Physics
- Reliability and Quality Control
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
- Epidemiology
- Physics
- Reliability and Quality Control
- Signal Detection
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
- Epidemiology
- Physics
- Reliability and Quality Control
- Signal Detection
- Social Networks
Scan statistics are used for detecting clusters of rare events.

Applications:

- Agricultural Sciences
- Astronomy
- Bioinformatics
- Biosurveillance
- Ecology and Environmental Sciences
- Epidemiology
- Physics
- Reliability and Quality Control
- Signal Detection
- Social Networks
- Telecommunication Sciences
References

References

References


References

References

References

References

More References


More References


More References


More References

More References


More References

More References

More References

Bonferroni-type Approximations and Inequalities


Theory and Methods

- **Bonferroni-type Approximations and Inequalities**

- **Finite Markov Chain Embedding**
Large Deviation Approximations


Theory and Methods

- **Large Deviation Approximations**


- **Martingale Formulation via Betting Systems**

Perfect Simulation Algorithms


Poisson and Compound-Poisson Approximations


Perfect Simulation Algorithms


Poisson and Compound-Poisson Approximations


Theory and Methods

- **Product-type Approximations and Inequalities**


Theory and Methods

- **Product-type Approximations and Inequalities**


- **Saddle-point Approximations**

Theory and Methods

- **FDR and FDC Methods**


- **GLR-type Tests via Simulation**


Theory and Methods

- **FDR and FDC Methods**


- **GLR-type Tests via Simulation**


Theory and Methods

- Linear Programming
- Monte Carlo Methods
- Order Statistics and Spacings
- Symbolic Computing
Probability Inequalities

Let $X_1, \ldots, X_N, \ldots$ be a sequence of iid normal observations with mean $\mu$ and variance $\sigma^2$. Let $Y_{r,u} = \sum_{i=r}^{u} X_i$ for $u \geq r \geq 1$. For integers $2 \leq m < N$, where $m$ is the length of the sliding window and $N$ is the specified range of the monitored process, define the scan statistic

$$S_{m,N} = \max_{m \leq j \leq N} \{ Y_{j-m+1,j} \}.$$  

(1)
Let $X_1, ..., X_N, ...$ be a sequence of iid normal observations with mean $\mu$ and variance $\sigma^2$. Let $Y_{r,u} = \sum_{i=r}^{u} X_i$ for $u \geq r \geq 1$. For integers $2 \leq m < N$, where $m$ is the length of the sliding window and $N$ is the specified range of the monitored process, define the \textit{scan statistic}

$$S_{m,N} = \max_{m \leq j \leq N} \{ Y_{j-m+1,j} \}. \quad (1)$$

The sequence $\{ Y_{j-m+1,j}; m \leq j \leq N \}$, based on which the scan statistic is defined, contains $N - m + 1$ observations of moving sums of length $m$. 
Let $X_1, \ldots, X_N, \ldots$ be a sequence of iid normal observations with mean $\mu$ and variance $\sigma^2$. Let $Y_{r,u} = \sum_{i=r}^{u} X_i$ for $u \geq r \geq 1$. For integers $2 \leq m < N$, where $m$ is the length of the sliding window and $N$ is the specified range of the monitored process, define the scan statistic

$$S_{m,N} = \max_{m \leq j \leq N} \{ Y_{j-m+1,j} \}.$$  \hfill (1)

The sequence $\{ Y_{j-m+1,j}; m \leq j \leq N \}$, based on which the scan statistic is defined, contains $N - m + 1$ observations of moving sums of length $m$.

The random variables $\{ Y_{j-m+1,j}; m \leq j \leq N \}$ have a joint multivariate normal distribution with mean vector $(m\mu, \ldots, m\mu)'$ and variance and covariance matrix $\Sigma = \{ \sigma_{i,j} \}$, where $\sigma_{i,i} = m\sigma^2$, $\sigma_{i,j} = 0$, for $|j - i| \geq m$ and $\sigma_{i,j} = (m - k)\sigma^2$, for $|j - i| = k$, $1 \leq k \leq m - 1$. 

Joseph Glaz (University of Connecticut)
Let $X_1, \ldots, X_N, \ldots$ be a sequence of iid normal observations with mean $\mu$ and variance $\sigma^2$. Let $Y_{r,u} = \sum_{i=r}^{u} X_i$ for $u \geq r \geq 1$. For integers $2 \leq m < N$, where $m$ is the length of the sliding window and $N$ is the specified range of the monitored process, define the scan statistic

$$S_{m,N} = \max_{m \leq j \leq N} \{Y_{j-m+1,j}\}.$$  \hfill (1)

The sequence $\{Y_{j-m+1,j}; m \leq j \leq N\}$, based on which the scan statistic is defined, contains $N - m + 1$ observations of moving sums of length $m$.

The random variables $\{Y_{j-m+1,j}; m \leq j \leq N\}$ have a joint multivariate normal distribution with mean vector $(m\mu, \ldots, m\mu)'$ and variance and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, where $\sigma_{i,i} = m\sigma^2$, $\sigma_{i,j} = 0$, for $|j - i| \geq m$ and $\sigma_{i,j} = (m - k)\sigma^2$, for $|j - i| = k$, $1 \leq k \leq m - 1$.

For $2 \leq m \leq N$ and $-\infty < t < \infty$, let

$$G_{m,t}(N) = P(\{Y_{1,m} < t, Y_{2,m+1} < t, \ldots, Y_{N-m+1,N} < t\}).$$ \hfill (2)
The distribution of the scan statistic $S_{m,N}$ is given by

$$P(S_{m,N} < t) = G_{m,t}(N).$$

(3)
The distribution of the scan statistic $S_{m,N}$ is given by

$$P(S_{m,N} < t) = G_{m,t}(N).$$ \hspace{1cm} (3)

The probability that the scan statistic exceeds level $t$ is given by

$$P(S_{m,N} \geq t) = 1 - G_{m,t}(N).$$ \hspace{1cm} (4)
The distribution of the scan statistic $S_{m,N}$ is given by

$$P(S_{m,N} < t) = G_{m,t}(N).$$

The probability that the scan statistic exceeds level $t$ is given by

$$P(S_{m,N} \geq t) = 1 - G_{m,t}(N).$$

This scan statistic can be used in detecting a local change in the process mean within a sequence of $N$ observations via testing the null hypothesis of randomness, $H_0$, that assumes $X_i, 1 \leq i \leq N$, are iid normal random variables with mean $\mu_0$ and variance $\sigma^2$. 
Probability Inequalities

- The distribution of the scan statistic $S_{m,N}$ is given by
  \[ P(S_{m,N} < t) = G_{m,t}(N). \] (3)
- The probability that the scan statistic exceeds level $t$ is given by
  \[ P(S_{m,N} \geq t) = 1 - G_{m,t}(N). \] (4)
- This scan statistic can be used in detecting a local change in the process mean within a sequence of $N$ observations via testing the null hypothesis of randomness, $H_0$, that assumes $X_i, 1 \leq i \leq N$, are iid normal random variables with mean $\mu_0$ and variance $\sigma^2$.
- For the alternative hypothesis, $H_1$, of a local change in $\mu$, one often specifies a segment of $m$ consecutive observations
  \[ R(i_0, m) = \{i_0, i_0 + 1, \ldots, i_0 + m - 1\}, \]
  where $1 \leq i_0 \leq N - m + 1$ is unknown and $2 \leq m \leq N/4$ is the window length. We first discuss the case when $m$ is known.
Under $H_1$, for any $i_0 \leq i \leq i_0 + m - 1$, $X_i$ has a normal distribution with mean $\mu_1$ and variance $\sigma^2$, where $\mu_1 > \mu_0$. For $i \notin R(i_0, m)$, $X_i$'s are distributed according to the distribution specified by the null hypothesis.
Under $H_1$, for any $i_0 \leq i \leq i_0 + m - 1$, $X_i$ has a normal distribution with mean $\mu_1$ and variance $\sigma^2$, where $\mu_1 > \mu_0$. For $i \notin R(i_0, m)$, $X_i$'s are distributed according to the distribution specified by the null hypothesis.

Let $X_1, \ldots, X_N, \ldots$ be iid continuous random variables with mean $\mu$ and variance $\sigma^2$. The following inequalities are used in approximating the distribution of the scan statistic:

Theorem (Glaz, Naus and Wang 2012) For integers $i, m \geq 2$, $L_1, G(N) - G(im) + G(Lm - 1) - G((L + 1)m - 1)iM$, $N(i - L)m$, (5) $G(N) - G(im)f[1 - G((L + 1)m - 1)gNim]$ for $N((i - L)m)$. 

Joseph Glaz (University of Connecticut) Scan Statistics for Normal Data 24/10 19 / 34
Probability Inequalities

- Under $H_1$, for any $i_0 \leq i \leq i_0 + m - 1$, $X_i$ has a normal distribution with mean $\mu_1$ and variance $\sigma^2$, where $\mu_1 > \mu_0$. For $i \notin R(i_0, m)$, $X_i$'s are distributed according to the distribution specified by the null hypothesis.

- Let $X_1, \ldots, X_N, \ldots$ be iid continuous random variables with mean $\mu$ and variance $\sigma^2$. The following inequalities are used in approximating the distribution of the scan statistic:

**Theorem (Glaz, Naus and Wang 2012)** For integers $i, m \geq 2, L \geq 1$,

\[
G(N) \geq \frac{G(im)}{1 + \frac{G(Lm-1) - G(Lm)}{G((L+1)m-1)}}^{M-im}, \quad N \geq (i \lor L)m, \quad (5)
\]

\[
G(N) \leq G(im) \{1 - [G((L + 1)m - 1) - G((L + 1)m)]\}^{N-im},
\]

for $N \geq (i \lor (L + 1))m$. 
The following inequalities are used for approximating the distribution of the scan statistic for normal data.
The following inequalities are used for approximating the distribution of the scan statistic for normal data.

For choices of $i$ and $L$ in theorem above and $G(3m - 1) \geq G(2m - 1)G(2m)$ we get:

$$G(N) \geq \frac{G(2m)}{\left[1 + \frac{G(2m-1)-G(2m)}{G(2m-1)G(2m)}\right]^{N-2m}}, \quad N \geq 2m, \quad (6)$$
The following inequalities are used for approximating the distribution of the scan statistic for normal data.

For choices of $i$ and $L$ in theorem above and $G(3m - 1) \geq G(2m - 1)G(2m)$ we get:

$$G(N) \geq \frac{G(2m)}{\left[1 + \frac{G(2m-1) - G(2m)}{G(2m-1)G(2m)}\right]^{N-2m}}, \quad N \geq 2m, \quad (6)$$

$$G(N) \leq G(2m)\{1 - [G(2m - 1) - G(2m)]\}^{N-2m}, \quad N \geq 2m. \quad (7)$$
The following inequalities are used for approximating the distribution of the scan statistic for normal data.

For choices of $i$ and $L$ in theorem above and $G(3m - 1) \geq G(2m - 1)G(2m)$ we get:

$$G(N) \geq \frac{G(2m)}{\left[1 + \frac{G(2m-1) - G(2m)}{G(2m-1)G(2m)}\right]^{N-2m}}, \quad N \geq 2m, \quad (6)$$

$$G(N) \leq G(2m)\{1 - [G(2m - 1) - G(2m)]\}^{N-2m}, \quad N \geq 2m. \quad (7)$$

We expect these bounds to be tight for a large value of $t$, since they converge as $G(2m) \to 1$ and $G(2m - 1) - G(2m) \to 0$, which holds as $t \to \infty$. 

The following inequalities are used for approximating the distribution of the scan statistic for normal data. For choices of $i$ and $L$ in theorem above and $G(3m - 1) \geq G(2m - 1)G(2m)$ we get:

$$G(N) \geq \frac{G(2m)}{1 + \frac{G(2m-1) - G(2m)}{G(2m-1)G(2m)}}^{N-2m}, \quad N \geq 2m, \quad (6)$$

$$G(N) \leq G(2m)\{1 - [G(2m-1) - G(2m)]\}^{N-2m}, \quad N \geq 2m. \quad (7)$$

We expect these bounds to be tight for a large value of $t$, since they converge as $G(2m) \to 1$ and $G(2m - 1) - G(2m) \to 0$, which holds as $t \to \infty$.

In Glaz, Naus and Wang (2012), inequalities for expected values and variances of a stopping time for moving sums are evaluated via the R algorithms in Genz and Bretz (2009).
A Markov-type approximation for $G(N)$ based on a method introduced in Naus (1982). Let $N = Km + \nu$, where $K \geq 3$, $m \geq 2$ and $0 \leq \nu \leq m - 1$ are integers. Then, for $2 \leq L \leq H - 1$

$$G(N) = P \left\{ \max_{m \leq k \leq N} Y_{k-m+1,k} < t \right\} = P \left( \bigcap_{i=1}^{K} E_j \right)$$

$$= P \left( \bigcap_{i=1}^{L-1} E_i \right) \prod_{j=L}^{K} P \left( E_j \mid \bigcap_{h=1}^{j-1} E_h \right), \quad (8)$$

where for $1 \leq j \leq K - 1$

$$E_j = \left( \max_{jm \leq k \leq (j+1)m} Y_{k-m+1,k} < t \right),$$

which can be interpreted as the event of no exceedance of level $t$ within a block of $m + 1$ consecutive partial sums of length $m$, and

$$E_K = \left( \max_{Km \leq k \leq Km+\nu} Y_{k-m+1,k} < t \right).$$
Product-type approximations

- By conditioning on the most recent past $L \geq 2$ events $E_j$, in (8) we get the following approximation for $G(M)$:

$$G(N) \approx P \left( \bigcap_{i=1}^{L-1} E_i \right) \left[ \prod_{j=L}^{K-1} P \left( E_j \bigg| \bigcap_{h=j-L+1}^{j-1} E_h \right) \right] P \left( E_K \bigg| \bigcap_{p=K-L+1}^{K-1} E_p \right)$$

$$= P \left( \bigcap_{i=1}^{L} E_i \right) \left[ \prod_{j=L+1}^{K-1} \frac{P \left( \bigcap_{h=j-L+1}^{j-1} E_h \right)}{P \left( \bigcap_{h=j-L+1}^{j-1} E_h \right)} \right] \frac{P \left( \bigcap_{p=K-L+1}^{K-1} E_p \right)}{P \left( \bigcap_{p=K-L+1}^{K-1} E_p \right)}$$

$$= G((L+1)m) \left[ \frac{G((L+1)m)}{G(Lm)} \right]^{K-L-1} \frac{G(Lm+\nu)}{G(Lm)}.$$  \hspace{1cm} (9)

- For $N = Km$ and $L = 2$ the above approximation reduces to

$$G(N) \approx G(3m) \left[ \frac{G(3m)}{G(2m)} \right]^{K-3}.$$  \hspace{1cm} (10)
Let $X_1, \ldots, X_N, \ldots$ be iid continuous random variables with mean $\mu$ and variance $\sigma^2$.

For $m \geq 2, j \geq 1$, let

$$q_j = P(Y_{j+1} \leq t | Y_{i+m-1} \leq t; 1 \leq i \leq j).$$

**Theorem** (Glaz and Johnson 1988): If $0 < P(X_1 \leq t/m) < 1$, then

$$\lim_{j \to \infty} q_j = q,$$

where $0 < q < 1$.

The proof of the theorem is based on the R-theory of Markov chains. One can show that for $m = 2$ the $q_j$'s oscillate about $q$. This property does not extend for $m \geq 3$, even though numerically one observes an oscillatory pattern of convergence of $q_j$ to $q$. 
Haiman (1999 and 2007) derived accurate approximations for $G(M)$ for iid discrete random variables. These approximations are valid as well for iid continuous random variables.

A nice feature of these approximations is that a sharp error bound can be easily evaluated.

For the problem at hand, for $N \geq 3m$, the following approximation for $G(N)$ is obtained from Haiman (2007, Corollary 2):

$$G(N) \approx \frac{2G(2m) - G(3m)}{\left[1 + G(2m) - G(3m) + 2(G(2m) - G(3m))^2\right]^{N/m}}, \quad (11)$$

with an error bound of approximately

$$3.3[1 - G(2m)]^2 N/m. \quad (12)$$
A Multiple Window Scan Statistics

Let \(2 \leq m_1 < m_2 < \ldots < m_n\) be a given sequence of window lengths associated with scan statistics \(S_{m_1}, \ldots, S_{m_n}\), respectively.
Let $2 \leq m_1 < m_2 < \ldots < m_n$ be a given sequence of window lengths associated with scan statistics $S_{m_1}, \ldots, S_{m_n}$, respectively.

Since the size of the rectangular window $m$ is unknown, for testing $H_0$ vs $H_1$ we propose the following test statistic:

$$P_{\text{min}} = \min\{p_j; 1 \leq j \leq n\},$$

the minimum $P$-value statistic, which is based on $n$ fixed window size scan statistics: $S_{m_1}, \ldots, S_{m_n}$, where $2 \leq m_j < m_{j+1} \leq N - 1$, $1 \leq j \leq n - 1$, and $p_j = P(S_m \geq k_j)$, is the observed p-value.
Let $2 \leq m_1 < m_2 < \ldots < m_n$ be a given sequence of window lengths associated with scan statistics $S_{m_1}, \ldots, S_{m_n}$, respectively.

Since the size of the rectangular window $m$ is unknown, for testing $H_0$ vs $H_1$ we propose the following test statistic:

$$P_{\text{min}} = \min\{p_j; 1 \leq j \leq n\},$$

(13)

the minimum $P$-value statistic, which is based on $n$ fixed window size scan statistics: $S_{m_1}, \ldots, S_{m_n}$, where $2 \leq m_j < m_{j+1} \leq N - 1$, $1 \leq j \leq n - 1$, and $p_j = P(S_m \geq k_j)$, is the observed p-value.

A simulation algorithm is used to implement this multiple window scan statistic and evaluate its power.
For $1 \leq j \leq n$, let $t_j$ be the observed value of $S_{m_j}$ and $p_j = P(S_{m_j} \geq t_j \mid H_0)$ the associated p-value. Since the exact distribution for the $P_{\min}$ statistic is unknown, for a given significant level $\alpha$, the critical value $p_{\alpha}$,

$$P_{H_0}(P_{\min} \leq p_{\alpha}) = \alpha,$$

has to be evaluated via simulation.
For $1 \leq j \leq n$, let $t_j$ be the observed value of $S_{m_j}$ and $p_j = P(S_{m_j} \geq t_j | H_0)$ the associated p-value. Since the exact distribution for the $P_{\text{min}}$ statistic is unknown, for a given significant level $\alpha$, the critical value $p_\alpha$, 

$$P_{H_0}(P_{\text{min}} \leq p_\alpha) = \alpha,$$

has to be evaluated via simulation.

In each run of the simulation, we generate $N$ observations under the null hypothesis. Then we scan the whole region with multiple moving windows of sizes $m_1, m_2, \ldots$ and $m_n$, and record the observed values of the fixed window scan statistics, $S_{m_1}, \ldots, S_{m_n}$, denoted by $t_1, t_2, \ldots, t_n$, respectively.
For $1 \leq j \leq n$, let $t_j$ be the observed value of $S_{m_j}$ and
$p_j = P(S_{m_j} \geq t_j \mid H_0)$ the associated p-value. Since the exact
distribution for the $P_{min}$ statistic is unknown, for a given significant
level $\alpha$, the critical value $p_\alpha$,

$$P_{H_0}(P_{min} \leq p_\alpha) = \alpha,$$

has to be evaluated via simulation.

In each run of the simulation, we generate $N$ observations under the
null hypothesis. Then we scan the whole region with multiple moving
windows of sizes $m_1, m_2, \ldots$ and $m_n$, and record the observed values
of the fixed window scan statistics, $S_{m_1}, \ldots, S_{m_n}$, denoted by
$t_1, t_2, \ldots, t_n$, respectively.

Then, a Monte-Carlo R algorithm is employed to evaluate the
observed p values: $p_j = P(S_{m_j} \geq t_j \mid H_0), 1 \leq j \leq n.$
A Multiple Window Scan Statistic

- For $1 \leq j \leq n$, let $t_j$ be the observed value of $S_{m_j}$ and $p_j = P(S_{m_j} \geq t_j \mid H_0)$ the associated p-value. Since the exact distribution for the $P_{\text{min}}$ statistic is unknown, for a given significant level $\alpha$, the critical value $p_\alpha$,

$$P_{H_0}(P_{\text{min}} \leq p_\alpha) = \alpha,$$

has to be evaluated via simulation.

- In each run of the simulation, we generate $N$ observations under the null hypothesis. Then we scan the whole region with multiple moving windows of sizes $m_1, m_2, \ldots$ and $m_n$, and record the observed values of the fixed window scan statistics, $S_{m_1}, \ldots, S_{m_n}$, denoted by $t_1, t_2, \ldots, t_n$, respectively.

- Then, a Monte-Carlo R algorithm is employed to evaluate the observed p values: $p_j = P(S_{m_j} \geq t_j \mid H_0), 1 \leq j \leq n$.

- The minimum of value of these p values is recorded and this process is repeated 10,000 times. Based on that, an approximate $\alpha \times 100$ percentile of the distribution of $P_{\text{min}}^{(1)}$ statistic is obtained.
Numerical Results

- Inequalities and approximations for a fixed window scan statistic for normal data with $\mu = 0$ and $\sigma^2 = 1$, $N = 1000$, $m = 50$:

<table>
<thead>
<tr>
<th>$t$</th>
<th>20</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LB$</td>
<td>.2530</td>
<td>.0729</td>
<td>.0479</td>
<td>.0388</td>
<td>.0221</td>
<td>.0121</td>
<td>.0074</td>
</tr>
<tr>
<td>$A1$</td>
<td>.2601</td>
<td>.0851</td>
<td>.0551</td>
<td>.0350</td>
<td>.0216</td>
<td>.0130</td>
<td>.0077</td>
</tr>
<tr>
<td>$A2$</td>
<td>.2587</td>
<td>.0847</td>
<td>.0551</td>
<td>.0349</td>
<td>.0214</td>
<td>.0132</td>
<td>.0077</td>
</tr>
<tr>
<td>$EB$</td>
<td>1.84–3</td>
<td>7.39–4</td>
<td>2.84–4</td>
<td>1.05–4</td>
<td>3.63–5</td>
<td>1.21–5</td>
<td></td>
</tr>
<tr>
<td>$UB$</td>
<td>.2607</td>
<td>.0886</td>
<td>.0613</td>
<td>.0390</td>
<td>.0252</td>
<td>.0121</td>
<td>.0085</td>
</tr>
</tbody>
</table>
Numerical Results

- Power study to evaluate the performance of the multiple window scan statistic: for normal data, $H_0 : \mu = 0, \sigma^2 = 1, N = 250$
- $\alpha = \Pr \text{ Type I Error}$, $\mu_1$ = the mean under the alternative in a subsequence of $n$ observations.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu_1$</th>
<th>$P_{\text{min}}$</th>
<th>$S_5$</th>
<th>$S_{10}$</th>
<th>$S_{15}$</th>
<th>$S_{20}$</th>
<th>$S_{25}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.5</td>
<td>.091</td>
<td>.074</td>
<td>.100</td>
<td>.076</td>
<td>.061</td>
<td>.067</td>
</tr>
<tr>
<td>1</td>
<td>.424</td>
<td>.339</td>
<td>.468</td>
<td>.317</td>
<td>.228</td>
<td>.189</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>.909</td>
<td>.828</td>
<td>.926</td>
<td>.770</td>
<td>.589</td>
<td>.465</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>.5</td>
<td>.150</td>
<td>.116</td>
<td>.149</td>
<td>.155</td>
<td>.119</td>
<td>.099</td>
</tr>
<tr>
<td>1</td>
<td>.742</td>
<td>.506</td>
<td>.695</td>
<td>.773</td>
<td>.618</td>
<td>.505</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>.993</td>
<td>.932</td>
<td>.987</td>
<td>.994</td>
<td>.971</td>
<td>.921</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>.5</td>
<td>.247</td>
<td>.153</td>
<td>.209</td>
<td>.263</td>
<td>.264</td>
<td>.189</td>
</tr>
<tr>
<td>1</td>
<td>.895</td>
<td>.602</td>
<td>.819</td>
<td>.878</td>
<td>.913</td>
<td>.834</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.0</td>
<td>.981</td>
<td>.997</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

$\alpha = .05 \quad .045 \quad .048 \quad .047 \quad .046 \quad .063$
Let $X_1, \ldots, X_M$ be a sequence of observations from an AR(1) process, $X_t = \theta X_{t-1} + \omega_t$, where $\omega_t$ is a Gaussian white noise with mean $\mu = 0$ and variance $\sigma^2 = 1$. Since $X_t$'s follow a multivariate normal distribution, $\{Y_{i-m+1,i}; m \leq i \leq M\}$ have a multivariate normal distribution with zero mean vector and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, where $\sigma_{i,j} = \text{cov}(Y_{i,i+m-1}, Y_{j,j+m-1})$.

A routine derivation, yields the following covariance matrix:

$$
\sigma_{i,j} = \begin{cases} 
\frac{\theta}{(1-\theta)^4} (1 - \theta^{j+m-i})(1 - \theta^{i-j}) + \frac{\theta^{j+m-i+1}}{(1-\theta)^4} (1 - \theta^{i-j})^2 \\
+ \frac{j+m-i}{(1-\theta)^2} + \frac{2\theta}{(1-\theta)^3} [j + m - 1 - i - \frac{\theta}{1-\theta} (1 - \theta^{j+m-1-i})] \\
+ \frac{\theta}{(1-\theta)^4} (1 - \theta^{i-j})(1 - \theta^{j+m-i}), i - j < m \\
\frac{1}{1-\theta^2} \{m + \frac{2\theta}{1-\theta} [m - 1 - \frac{\theta}{1-\theta} (1 - \theta^{m-1})]\}, i = j \\
\theta^{i-j-m+1} \frac{(1-\theta^m)^2}{(1-\theta)^2}, \text{ otherwise.}
\end{cases}
$$
Scan Statistics for Time Series Data

- Given the mean vector and covariance matrix, we can utilize the R algorithms by Genz and Bretz (2009) to approximate the distribution $G(M)$ for a fixed window scan statistic and the multiple window scan statistic $P_{\text{min}}$.

- For an AR(2) model, $X_t = \theta_1 X_{t-1} + \theta_2 X_{t-2} + \omega_t$, where $\omega_t$ is the Gaussian white noise with mean $\mu = 0$ and variance $\sigma^2 = 1$, the $X_t'$s follow a multivariate normal distribution with the following ACF:

$$
\gamma_h = \begin{cases} 
\frac{1-\theta_2}{1-\theta_1-\theta_2^2-\theta_2^2+\theta_2^3}, & \text{when } h = 0, \\
\gamma_0 \frac{\theta_1}{1-\theta_2}, & \text{when } h = 1, \\
\gamma_0 [\theta_1 \gamma_{h-1} + \theta_2 \gamma_{h-2}], & \text{when } h > 1.
\end{cases}
$$

- Then $\{Y_{i-m+1,i}; m \leq i \leq M\}$ a multivariate normal distribution with a mean vector of zeros and covariance matrix $\Sigma = \{\sigma_{i,j}\}$, which can be derived similarly as in the AR(1) process. The explicit form of the covariance matrix is omitted here for simplicity. Wang and Glaz (2013) investigated the performance of multiple window scan.
Numerical Results

- Power study to evaluate the performance of the multiple window scan statistic $P_{\text{min}}$: for AR(1) data, $\theta = .1$, $N = 1500$

- $\alpha = \Pr \text{ Type I Error}$, $\mu_1 = \text{the mean under the alternative in a subsequence of } n \text{ observations in the white noise component.}$

\begin{tabular}{cccccccc}
\hline
$n$ & $\mu_1$ & $P_{\text{min}}$ & $S_5$ & $S_{10}$ & $S_{15}$ & $S_{20}$ & $S_{25}$ \\
\hline
10 & .5 & .076 & .075 & .063 & .056 & .062 & .05 \\
1 & .292 & .211 & .337 & .172 & .124 & .101 & \\
1.5 & .797 & .649 & .841 & .554 & .388 & .285 & \\
15 & .5 & .103 & .072 & .087 & .093 & .085 & .074 \\
1 & .532 & .273 & .494 & .594 & .407 & .297 & \\
1.5 & .973 & .810 & .960 & .989 & .915 & .800 & \\
20 & .5 & .115 & .068 & .091 & .113 & .120 & .100 \\
1 & .762 & .378 & .615 & .759 & .818 & .683 & \\
1.5 & 1.0 & .905 & .992 & 1.0 & 1.0 & .992 & \\
\hline
$\alpha$ & .050 & .037 & .052 & .052 & .051 & .052 & \\
\hline
\end{tabular}
This data set consists of 310 hourly uncontrolled viscosity readings of a chemical process. This data set has been modeled via an AR(1) process in Box and Jenkins (1978), with estimated parameters: $\theta = 0.87$, and $\sigma^2 = 0.09$. 

To evaluate the performance of the multiple window scan statistic, we introduced a change in the Gaussian white noise component at a random location. We employed a similar algorithm to the one outlined above to perform a power study that is presented in the table below. A simulation with 10,000 trial has been used to simulate the power. The multiple window scan statistic outperformed the fixed window scan statistics, with an incorrectly specified window size where a change in mean has occurred. A discrepancy in some of the results could have resulted from the model lack of fit.
This data set consists of 310 hourly uncontrolled viscosity readings of a chemical process. This data set has been modeled via an AR(1) process in Box and Jenkins (1978), with estimated parameters: \( \theta = 0.87 \), and \( \sigma^2 = 0.09 \).

To evaluate the performance of the multiple window scan statistic, we introduced a change in the Gaussian white noise component at a random location. We employed a similar algorithm to the one outlined above to perform a power study that is presented in the table below. A simulation with 10,000 trial has been used to simulate the power.
This data set consists of 310 hourly uncontrolled viscosity readings of a chemical process. This data set has been modeled via an AR(1) process in Box and Jenkins (1978), with estimated parameters: $\theta = 0.87$, and $\sigma^2 = 0.09$.

To evaluate the performance of the multiple window scan statistic, we introduced a change in the Gaussian white noise component at a random location. We employed a similar algorithm to the one outlined above to perform a power study that is presented in the table below. A simulation with 10,000 trial has been used to simulate the power.

The multiple window scan statistic outperformed the fixed window scan statistics, with an incorrectly specified window size where a change in mean has occurred. A discrepancy in some of the results could have resulted from the model lack of fit.
Numerical Results

- Series D data set from Box and Jenkins (1978).
Numerical Results

- Series D data set from Box and Jenkins (1978).
- $\alpha = \Pr \text{ Type I Error, } \mu_1 = \text{the mean under the alternative in a subsequence of } n \text{ observations in the white noise component.}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\mu_1$</th>
<th>$P_{\min}$</th>
<th>$S_5$</th>
<th>$S_{10}$</th>
<th>$S_{15}$</th>
<th>$S_{20}$</th>
<th>$S_{25}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.15</td>
<td>.142</td>
<td>.170</td>
<td>.260</td>
<td>.035</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>.584</td>
<td>.588</td>
<td>.628</td>
<td>.330</td>
<td>.100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.703</td>
<td>.731</td>
<td>.710</td>
<td>.641</td>
<td>.403</td>
<td>.262</td>
</tr>
<tr>
<td>15</td>
<td>.15</td>
<td>.394</td>
<td>.234</td>
<td>.379</td>
<td>.473</td>
<td>.226</td>
<td>.161</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>.704</td>
<td>.725</td>
<td>.700</td>
<td>.664</td>
<td>.622</td>
<td>.520</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.841</td>
<td>.845</td>
<td>.834</td>
<td>.864</td>
<td>.756</td>
<td>.731</td>
</tr>
<tr>
<td>20</td>
<td>.15</td>
<td>.625</td>
<td>.324</td>
<td>.499</td>
<td>.574</td>
<td>.617</td>
<td>.500</td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>.778</td>
<td>.787</td>
<td>.774</td>
<td>.750</td>
<td>.753</td>
<td>.737</td>
</tr>
<tr>
<td></td>
<td>.25</td>
<td>.932</td>
<td>.889</td>
<td>.914</td>
<td>.910</td>
<td>.942</td>
<td>.805</td>
</tr>
<tr>
<td>$\alpha$</td>
<td></td>
<td>.051</td>
<td>.040</td>
<td>.054</td>
<td>.049</td>
<td>.059</td>
<td>.040</td>
</tr>
</tbody>
</table>
Summary and Future Work

- Two dimensional continuous-type data sets
- Scan statistics for graphs
- Non-homogeneous processes
- Three dimensional scan statistics
- Conditional-type scan statistics