SCAN STATISTICS WITH WEIGHTED OBSERVATIONS

Hock Peng Chan* and Nancy Ruonan Zhang†

Abstract

We examine scan statistics for one dimensional marked Poisson processes. Such statistics tabulate the maximum weighted count of event occurrences within a window of pre-determined width over all windows within an observed interval. We derive analytical formulas as well as give an importance sampling method for approximating the tail probabilities of scan statistics.

As high throughput genomic sequencing led to the availability of massive amounts of biomolecular sequence data, it is often of interest to search long DNA or protein sequences for local regions that are enriched for a certain characteristic. Thus, scan statistics have become a useful tool in modern computational biology. We illustrate the application of our p-value approximations with such examples.

KEY WORDS: Change of measure, DNA sequence, importance sampling, large deviations, marked Poisson process, scan statistics.

*Department of Statistics and Applied Probability, National University of Singapore, e-mail: stachp@nus.edu.sg
†Department of Statistics, University of California, Berkeley, e-mail: nzhang@stat.berkeley.edu
1 Introduction

Scan statistics are used for detecting the presence of an unusually large cluster of events ordered by either time or location. A scanning window is moved along the time (or location) interval and the maximum number of events that is captured by this window at some point in time is recorded. Due to the multiple comparisons effect from maximizing over all possible windows, most of which are non-overlapping, scan statistics are often larger than expected from mere intuition when there is in fact no clustering present. Glas, Naus and Wallenstein (2001, p29-31) illustrated this point with interesting examples and motivated the need for precise p-value computations for scan statistics. They documented the development of scan statistic tail approximations and bounds over the past forty years, and provided detailed derivations for these estimates. For more detailed discussions, see Naus (1965, 1982), Cressie (1980), Glaz (1989), Glaz and Naus (1991) and Loader (1991).

One particular area in which scan statistics have been useful in recent years is the computational analysis of DNA and protein sequences. For example, Lifanov et al. (2003) scanned DNA sequences for clusters of transcription factor binding sites in order to locate genes that relate to specific biological processes. They used position weight matrices to score words for similarity to a given transcription factor pattern, and used a cut-off value for the word score to determine locations of occurrence of the pattern. Rajewsky et al. (2002) considered a similar problem with the exception that the total score of all words in a window exceeding the cut-off was used to compute the scan statistics instead of the number of words exceeding the cut-off. Currently available p-value approximations for scan statistics of point processes treat only the case of 0-1 processes that are applicable to Lifanov et al. (2003), but cannot handle directly
a scoring or weighting scheme used in Rajewsky et al. (2002).

In Section 2, we provide p-value approximations for scan statistics of marked Poisson processes. Our approximations are applicable to general scoring schemes used in computational biology. A novel feature of our formula is an overshoot correction term that disappears in the special case of 0-1 processes. We apply our p-value approximations to various problems in computational biology in Section 3. In Section 4, we use Monte Carlo methods to check the analytical p-value approximations. Since some of the p-values considered are very small, which renders direct Monte Carlo inefficient, we propose an importance sampling scheme to provide more accurate estimates. We end this paper in Section 5 with a discussion on computational issues and choice of scoring functions.

2 Theoretical results

Let \( N \) be a Poisson process on \((0, n]\) with constant rate \( \lambda > 0 \). Let \( X_1, X_2, \ldots \) be independent and identically distributed (i.i.d.) random variables with cumulative distribution function \( F \) and independent of \( N \) such that \( \mu := EX_1 > 0 \). We say that \( F \) is arithmetic if the support of \( F \) lies on \( \pm \eta, \pm 2\eta, \ldots \) for some \( \eta > 0 \). The largest \( \eta \) with this property is known as the span of \( F \) (see Feller (1971) Section 5.2). Let \( S_k = X_1 + \cdots + X_k \). Thus for any window \((t, t + u]\), \( S_{N(t+u)} - S_N(t) \) is the weighted count of the point process in that window. We define

\[
M_{n,u} = \sup_{0 \leq t \leq n-u} [S_{N(t+u)} - S_N(t)],
\]

where \( u \in (0, n) \) is a pre-determined width of the sliding window. The scan statistic

\[
\sup_{0 \leq t \leq n-u} [N(t + u) - N(t)]
\]

that has been widely studied is a special case of (1) when \( F \) is degenerate at 1. In (1), the occurrence of the \( i \)th jump in \( N \) is weighted by a score \( X_i \) and hence we call \( M_{n,u} \) a weighted scan statistic. Similarly, we call (2) an unweighted scan statistic.
In Theorem 1 below, we will give an asymptotic approximation for the tail probability of the weighted scan statistic. Before stating the theorem, we will need to define some constants. Assume that the moment generating function of $F$, say $K(\theta) = E(e^{\theta X_1})$, is finite for some $\theta > 0$. Let $K'(\theta) = \frac{d}{d\theta} K(\theta) = E(X_1 e^{\theta X_1})$ and $K''(\theta) = \frac{d^2}{d\theta^2} K(\theta) = E(X_1^2 e^{\theta X_1})$. Given $x > \lambda \mu$, let $\theta_x > 0$ and $\alpha_x > \lambda$ be the unique constants that satisfy

$$K'(\theta_x) = (x/\lambda) \quad \text{and} \quad \alpha_x = \lambda K(\theta_x).$$

(3)

We also define the large deviation rate function

$$I(x) = -(\alpha_x - \lambda) + \theta_x x.$$

(4)

To make the notation simple, we shall assume here that either $F$ is continuous with density $f$ or discrete with probability mass function $f$. Embed $F$ in an exponential family of distribution functions $\{F_\theta\}$, where $F_\theta$ has a density or probability mass function $f_\theta$ satisfying

$$f_\theta(y) = e^{\theta y} f(y)/K(\theta) \quad \text{whenever} \ K(\theta) < \infty.$$  

(5)

Let $Y_1, Y_2, \ldots$ be i.i.d. random variables with the mixture density or probability mass function

$$g(y) = \left(\frac{\alpha_x}{\lambda + \alpha_x}\right) f_{\theta_x}(y) + \left(\frac{\lambda}{\lambda + \alpha_x}\right) f(-y),$$

(6)

and let $R_k = Y_1 + \cdots + Y_k$. We define the overshoot constant

$$\nu_x = \lim_{b \to \infty} E[e^{-\theta_x (R_{\tau_b} - b)}], \quad \text{where} \ \tau_b = \inf\{k \geq 1 : R_k \geq b\},$$

(7)

with the convention that $b$ in (7) is a multiple of $\eta$ if $F$ is arithmetic with span $\eta$. For more details on the overshoot constant, for example its existence and closed form expressions, see Siegmund (1985, Chapter 8), Tu and Siegmund (1999) and Storey and Siegmund (2001).

For any $a \in \mathbb{R}$ we let $\lfloor a \rfloor$ denote the greatest integer not exceeding $a$. For two sequences $a_n, b_n$, by $a_n \sim b_n$ we mean that $\lim_{n \to \infty} (a_n/b_n) = 1$. 

3
Theorem 1. Let $v > \lambda \mu$ be fixed. If $F$ is arithmetic with span $\eta > 0$, let $x(= x_u) = \eta u^{-1} \lfloor uv/\eta \rfloor$. Thus $ux$ is the largest multiple of $\eta$ not exceeding $uv$. For $F$ that is non-arithmetic, let $x = v$. Let $u \to \infty$ as $n \to \infty$ such that $(n - u) \to \infty$. Then,

$$P\{M_{n,u} \geq ux\} \sim 1 - \exp\left\{-\frac{(n - u)\nu_x e^{-uI(x)(x - \lambda \mu)}}{\sqrt{2\pi u \lambda K''(\theta_x)}}\right\}.$$ (8)

It follows from (3), (6) and (7) that for the unweighted scan statistic in which $F$ is degenerate at 1,

$$K(\theta_x) = K'(\theta_x) = K''(\theta_x) = e^{\theta_x}, \quad \alpha_x = x, \quad \theta_x = \log(x/\lambda) \text{ and } \nu_x = 1.$$ (9)

Hence Theorem 1 reduces to the following for this special case:

**Corollary 1.** Let $v > \lambda$ be fixed and define $x = u^{-1} \lfloor uv \rfloor$. Let $u \to \infty$ as $n \to \infty$ such that $(n - u) \to \infty$. Then,

$$P\left\{\sup_{0 \leq t \leq n-u} [N(t+u) - N(t)] \geq ux\right\} \sim 1 - \exp\left\{-\frac{(n - u)e^{u(x-\lambda)}(\lambda/x)^{ux}(x - \lambda)}{2\pi ux}\right\}.$$ (10)

**Remarks 1.** If $u \to \infty$ and $n/u \to \infty$ as $n \to \infty$, the relations (8) and (10) still hold if $\lambda$ is replaced by the consistent estimator $\hat{\lambda} = N(n)/n$. For problems in computational biology, $n$ often represents the length of a genome. Since genomes are very long and the purpose of using scan statistics is to target relatively short segments of the genome of length $u$ with “unusual” characteristics for further biological analysis, the required assumption is satisfied in practice. This allows us to estimate the Poisson rate parameter $\lambda$ from the data.

2. Frolov (2005) considers the class of stochastically continuous processes with independent increments for which the marked Poisson process that we consider here is a special case. Bounds on the tail probabilities of $M_{n,u}$ were obtained and applied, using Borel-Cantelli lemmas, to derive almost sure limits of $M_{n,u}$ as $u \to \infty$. The tail approximations in (8) are however much sharper than those in Frolov (2005).
3 Examples

Before getting into the examples, we first describe the computation of the weighted scan statistic. Let \( t_i \) denote the time of occurrence of the \( i \)th event of interest and \( X_i \) the score associated with this event. We construct the counting process \( N(t) = \sum_i 1\{t_i \leq t\} \), where \( 1 \) denotes the indicator function. Then for a given window-size \( u \), the weighted scan statistic

\[
M_{n,u} = \sup_{0 \leq t \leq n-u} \left( \sum_{i:t_i \leq t+u} X_i \right).
\]

(11)

For computational purposes, it suffices to evaluate the sum on the right hand side of (11) at \( t = t_i \) and \( t = t_i - u \) for all \( i \).

We shall next describe the computation of the p-value approximations in (8). If \( \lambda \) is unknown, we replace it by \( \hat{\lambda} = N(n)/n \). Unknown parameters of \( F \), for example \( \mu \), can also be estimated from the data. This will be elaborated in Example 2. Since \( K' \) is an increasing function, \( \theta_x \) can be computed via bijection methods using (3) if it cannot be obtained analytically. The constant \( K''(\theta_x) \) and large deviation rate \( I(x) \) can then be computed using (3) and (4).

It remains for us to compute the overshoot constant \( \nu_x \). Let \( \tau_+ = \inf\{ k \geq 1 : R_k > 0 \} \). By renewal theory, see for example Siegmund (1985) Corollary 8.33, it follows that if \( F \) is nonarithmetic, then

\[
\lim_{b \to \infty} P\{ R_{\tau_+} - b > y \} = (ER_{\tau_+})^{-1} \int_y^{\infty} P\{ R_{\tau_+} > z \} \, dz \text{ for } y \geq 0,
\]

(12)

and if \( F \) is arithmetic with span \( \eta \), then

\[
\lim_{b \to \infty, b \in \eta \mathbb{Z}} P\{ R_{\tau_+} - b = j\eta \} = \eta(ER_{\tau_+})^{-1} P\{ R_{\tau_+} \geq (j+1)\eta \} \text{ for } j = 0, 1, \ldots.
\]

(13)

In Example 2, we consider \( F \) geometric with mean \( \mu \). By (5), \( F_{\theta_x} \) is also geometric. Hence by (6) and the memoryless property of the geometric distribution, \( R_{\tau_+} \) is geometric with
distribution $F_{\theta_x}$. We can then use (7) and (13) to show that

$$\nu_x = \mu[1 - (1 - \mu^{-1})e^{\theta_x}].$$

(14)

In Examples 1 and 3, we check Corollary 1 against p-value approximations given in Naus (1982) for the unweighted scan statistics. For the weighted scan statistics in Example 2, we rely upon importance sampling to demonstrate the accuracy of (8).

**Example 1.** Biologists are interested in finding segments of the genome that contain a high concentration of palindromic patterns (PLP) as these segments are likely to be near an origin of replication of the virus (cf. Masse et al., 1992). The DNA alphabet has four letters A,T,C,G with A-T and C-G being complementary pairs on opposite strands of the DNA helix. Thus the complementary DNA sequence of GGATCC would be CCTAGG. The DNA sequence GGATCC is a PLP because its complement reads the same as itself backwards. We define the length of a palindrome to be the number of complementary pairs that it contains (i.e. the length of GGATCC is 3).

Denote by PLP* a PLP with length of at least 5 bp that is not nested inside another PLP. In Leung, Schactel and Yu (1994), the occurrence of PLP* in the Human cytomegalovirus (HCMV) genome was modeled as a Poisson process. The genome contains $n = 229,354$ base pairs and a total of $N(n) = 296$ PLP* were observed. The rate of the Poisson process was thus estimated by $\hat{\lambda} = N(n)/n = 296/229,354 = 0.00129$. The unweighted scan statistic for window size $u = 1000$ bp was computed and found to be equal to 10. The corresponding p-value of 0.00195 computed via Corollary 1 agrees quite well with the estimate of 0.00193 obtained using Naus (1982) and the Monte Carlo estimate of 0.0021 obtained in the next section (see Table 3).

**Example 2.** We continue on the problem of finding clusters of palindromic patterns in viral genomes. Instead of giving equal weightage to each PLP*, we now assign a score of
$X_i = \ell_i - 4$, where $\ell_i$ is the length of the $i$th PLP*. By definition, palindromes always have even lengths. We define the location $t_i$ of the $i$th PLP* to be the location of its left center. We then compute the weighted scan statistics $M_{n,u}$ with window length $u$ equal to 0.5% of the genome length, rounded off to the nearest 100 bases. We apply (8) with estimated Poisson rate $\hat{\lambda} = \frac{N(n)}{n}$ and geometric $F$ with estimated mean $\hat{\mu} = (1 - 2\hat{a}_A\hat{a}_T - 2\hat{a}_C\hat{a}_G)^{-1}$, where $(\hat{a}_A, \hat{a}_C, \hat{a}_G, \hat{a}_T)$ are the empirical probabilities of the four bases in the genome.

Chew, Choi and Leung (2005) also studied clustering of PLP* but used the score $X_i = \ell_i$ (or equivalently $X_i = \ell_i/5$), corresponding to a shifted geometric distribution for $X_i$. They did not provide any p-value approximations and instead of (1), they considered a scan statistic in which consecutive windows differ by half the window length, or more specifically, $\sup_{k \in \mathbb{Z}, 0 \leq k \leq (2n/u)-1}[S_{N(uk/2+u)} - S_{N(uk/2)}]$.

Figure 1 plots the unweighted and weighted scan statistics against genome location for three viruses: Cercopithecine herpesvirus 1 (CeHV1), Bovine herpesvirus 1 (BoHV1) and Bovine herpesvirus 5 (BoHV5). Origins of replication for these viruses that have been validated experimentally are also shown in the figure. To avoid excessive number of false positives when working with a large number of genomes, we choose a conservative p-value cutoff of 0.001 and use (8) to determine the threshold levels corresponding to this cutoff. Table 1 provides the computed scan statistics and their p-value approximations. We see from Figure 1 that using a length based weighting scheme improves the power for both CeHV1 and BoHV1. For BoHV5, significant clusters of palindromes are found in the vicinity of the replication origins. However, for this genome there are also many false positives.

Example 3. The location of the pattern GATC in a DNA sequence is known as a DAM site and is related to the repair and replication of DNA. An E. coli genome sequence has approximately $n = 4.7$ million bp with an approximate rate of $\lambda = 1.1/250$ occurrences of DAM sites per base pair. We are interested here in finding an unusually large number of
Figure 1: Comparison of weighted and unweighted scan statistics for 3 viral genomes. For all plots, horizontal axis is location in genome. The top plots show the locations and length of palindromes longer than 4. The middle plots show $u^{-1}[N(t + u/2) - N(t - u/2)]$ against $t$ in the unweighted case. The bottom plots show $u^{-1}(S_{N(t+u/2)} - S_{N(t-u/2)})$ against $t$ in the weighted case. Experimentally validated replication origins are indicated by triangles at the top of the plots. Dashed horizontal lines show thresholds for p-value of 0.001.
Table 1. Summary of information for the scan statistics of three viral genomes.

<table>
<thead>
<tr>
<th></th>
<th>($\hat{a}_A, \hat{a}_C, \hat{a}_G, \hat{a}_T$)</th>
<th>$n$</th>
<th>$N(n)$</th>
<th>$u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CeHV1</td>
<td>(0.13, 0.37, 0.38, 0.13)</td>
<td>156 789</td>
<td>580</td>
<td>800</td>
</tr>
<tr>
<td>BoHV1</td>
<td>(0.14, 0.36, 0.37, 0.14)</td>
<td>135 301</td>
<td>615</td>
<td>700</td>
</tr>
<tr>
<td>BoHV5</td>
<td>(0.12, 0.37, 0.38, 0.13)</td>
<td>138 390</td>
<td>714</td>
<td>700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Unweighted $M_{n,u}$</th>
<th>p-value</th>
<th>$F$ geometric $M_{n,u}$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CeHV1</td>
<td>18</td>
<td>$7.23 \times 10^{-6}$</td>
<td>116</td>
<td>0</td>
</tr>
<tr>
<td>BoHV1</td>
<td>17</td>
<td>$1.09 \times 10^{-4}$</td>
<td>32</td>
<td>$6.08 \times 10^{-5}$</td>
</tr>
<tr>
<td>BoHV5</td>
<td>15</td>
<td>$1.07 \times 10^{-2}$</td>
<td>33</td>
<td>$1.74 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 2. Estimation of $P\{M_{n,u} \geq ux\}$ (± standard error for direct Monte Carlo).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monte Carlo</td>
<td>Estimate (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.876±0.007</td>
<td>0.873</td>
<td>0.870</td>
<td>0.999</td>
</tr>
<tr>
<td>9</td>
<td>0.25±0.01</td>
<td>0.246</td>
<td>0.244</td>
<td>0.0987</td>
</tr>
<tr>
<td>10</td>
<td>0.034±0.004</td>
<td>0.0337</td>
<td>0.0334</td>
<td>0.011</td>
</tr>
</tbody>
</table>

DAM sites in a sliding window of length $u = 245$ bp. A formula equivalent to the Newell-Ikeda approximation was used in Karlin and Brendel (1992) to compute the p-value of the scan statistics. Glaz, Naus and Wallenstein (2001), however, suggested using Naus (1982). We compare these estimates with (10) and a Monte Carlo derived estimate involving 2000 simulation runs; see Table 2.
4 Importance sampling for small p-values

Let \( \mathbf{X}^{(i)} = \{ N^{(i)}, X_1^{(i)}, \ldots, X_{N^{(i)}(n)}^{(i)} \} \), 1 \leq i \leq B be \( B \) independent realizations of the underlying marked Poisson process with rate \( \lambda \) and weights following a distribution \( F \). Then

\[
\hat{p}_D = B^{-1} \sum_{i=1}^{B} \mathbf{1}_{\{ M_n^{(i)} \geq ux \}} \quad \text{and} \quad \text{s.e.}_D = [\hat{p}_D(1 - \hat{p}_D)/B]^{1/2}
\]  

(15)

are the unbiased direct Monte Carlo estimate of \( p := P\{M_{n,u} \geq ux\} \) and its consistent standard error estimate respectively. We provide below an alternative way of simulating the probability of \( \{ M_n^{(i)} \geq ux \} \) that attains substantial variance reduction when the probabilities are small.

Let \( \theta > 0 \) and \( \alpha_x > \lambda \) satisfy (3). For each 1 \( \leq i \leq B \), we do the following:

1. Generate \( t^{(i)} \) uniformly from \([0, n - u]\).

2. Generate \( N^{(i)} \) from a non-uniform Poisson process with rate \( \alpha_x \) on the interval \((t^{(i)}, t^{(i)} + u]\) and rate \( \lambda \) elsewhere on \((0, n]\).

3. Generate independent random variables \( X_1^{(i)}, \ldots, X_{N^{(i)}(n)}^{(i)} \) with \( X_j^{(i)} \) having distribution \( F_{\theta_x} \) for \( N^{(i)}(t^{(i)}) < j \leq N^{(i)}(t^{(i)} + u) \) and distribution \( F \) otherwise.

Let \( S_k^{(i)} = X_1^{(i)} + \cdots + X_k^{(i)} \). Then by (3) and (4), the likelihood ratio of \( \mathbf{X}^{(i)} \) between the underlying marked point process and the process generated via steps 1-3 above is

\[
L_i = \left[ (n - u)^{-1} \int_0^{n-u} e^{\theta_x(S_k^{(i)}(t+u) - S_k^{(i)}(t))/[N^{(i)}(t+u) - N^{(i)}(t)]} \right]^{-1}
\]

\[
\times e^{-(\alpha_x - \lambda)u(\alpha_x/\lambda)N^{(i)}(t+u) - N^{(i)}(t)} dt
\]

\[
= e^{-uI(x)}(n - u) / \left( \int_0^{n-u} e^{\theta_x(S_k^{(i)}(t+u) - S_k^{(i)}(t) - ux)} dt \right).
\]

(16)

The corresponding unbiased estimate of \( p \) and its respective standard error estimate are then

\[
\hat{p}_I = B^{-1} \sum_{i=1}^{B} L_i \mathbf{1}_{\{ M_n^{(i)} \geq ux \}} \quad \text{and} \quad \text{s.e.}_I = B^{-1} \left( \sum_{i=1}^{B} L_i^2 \mathbf{1}_{\{ M_n^{(i)} \geq ux \}} - B\hat{p}_I^2 \right)^{1/2}.
\]

(17)
For the unweighted scan statistics in which $F = F_{\theta_x}$ are degenerate at 1, step 3 above can be omitted. By (9) and (16),

$$L_i = (n - u)e^{(x - \lambda)u} / \left( \int_0^{n-u} (x/\lambda)^{N(t)}(t+u) - N(t) dt \right).$$

(18)

For the degenerate case, a similar importance sampling scheme was introduced in Naiman and Priebe (2001). Change of measure importance sampling schemes have also been used in sequential analysis (Siegmund, 1976 and Chan and Lai, 2000), change-point detection (Lai and Shan, 1999) and sequence alignments (Chan, 2003). The change of measure associated with the importance sampling scheme above also plays an important role in deriving the asymptotic expression of $p$ in Theorem 1.

Example 4. We check some of the analytical estimates of $p = P\{M_{n,u} \geq ux\}$ that was applied in Examples 1 and 2 via direct Monte Carlo and importance sampling with $B = 2000$ runs for each entry. In Table 3, we check the approximations for the HCMV genome with $n = 229\,354$, $u = 1000$ and $\lambda = 296/n = 0.00129$. Here we consider the unweighted scan statistics. In Table 4, we check the approximations for the BoHV1 genome with $n = 135\,301$, $u = 700$, $\lambda = 615/n = 0.00455$ and $F$ geometric with mean $\mu = (0.700)^{-1}$. In Table 5, we check the approximations with parameters taken from the BoHV5 genomes, with $n = 138\,390$, $u = 700$, $\lambda = 714/n = 0.00516$ and with $P\{X_1 = k\} = (0.3124)^{k-5}(0.6876)$ for $k \geq 5$. We note that for $F$ degenerate at 1, $\nu_x = 1$; for the geometric distribution, we compute $\nu_x = \mu[1 - (1 - \mu^{-1})e^{\theta_x}]$, see (14). We calculate the overshoot of the shifted geometric by recursive computation of the distribution of $R_{\min(t,\tau_+)}$ for $t = 1, 2, \ldots$ and then applying (7) and (13).

We see in Tables 3 to 5 substantial variance reduction for probabilities of the order $10^{-2}$ or $10^{-3}$ when importance sampling is used in place of direct Monte Carlo. For probabilities of order $10^{-4}$ or $10^{-5}$, direct Monte Carlo breaks down whereas the importance sampling algorithm still provides reliable estimates. The analytical estimates (8) and (10) agree very
Table 3. Estimation of $p \pm$ s.e. with $F$ degenerate at 1.

<table>
<thead>
<tr>
<th>$ux$</th>
<th>Direct</th>
<th>Importance</th>
<th>Analytical</th>
<th>Naus (1982)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monte Carlo</td>
<td>Sampling</td>
<td>Estimate (10)</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>$(1.5 \pm 0.3) \times 10^{-2}$</td>
<td>$(1.3 \pm 0.07) \times 10^{-2}$</td>
<td>$1.32 \times 10^{-2}$</td>
<td>$1.32 \times 10^{-2}$</td>
</tr>
<tr>
<td>10</td>
<td>$(1 \pm 1) \times 10^{-3}$</td>
<td>$(2.1 \pm 0.1) \times 10^{-3}$</td>
<td>$1.95 \times 10^{-3}$</td>
<td>$1.93 \times 10^{-3}$</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>$(2.2 \pm 0.2) \times 10^{-4}$</td>
<td>$2.53 \times 10^{-4}$</td>
<td>$2.53 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

Table 4. Estimation of $p \pm$ s.e. with $F$ geometric.

<table>
<thead>
<tr>
<th>$ux$</th>
<th>Direct</th>
<th>Importance</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monte Carlo</td>
<td>Sampling</td>
<td>Estimate (8)</td>
</tr>
<tr>
<td>24</td>
<td>$(1.0 \pm 0.2) \times 10^{-2}$</td>
<td>$(1.42 \pm 0.07) \times 10^{-2}$</td>
<td>$1.5 \times 10^{-2}$</td>
</tr>
<tr>
<td>26</td>
<td>$(4 \pm 1) \times 10^{-3}$</td>
<td>$(4.3 \pm 0.3) \times 10^{-3}$</td>
<td>$4.04 \times 10^{-3}$</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>$(1.00 \pm 0.07) \times 10^{-3}$</td>
<td>$1.04 \times 10^{-3}$</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>$(2.5 \pm 0.2) \times 10^{-4}$</td>
<td>$2.55 \times 10^{-4}$</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>$(5.3 \pm 0.4) \times 10^{-5}$</td>
<td>$6.08 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Table 5. Estimation of $p \pm$ s.e. with $F$ shifted geometric.

<table>
<thead>
<tr>
<th>$ux$</th>
<th>Direct</th>
<th>Importance</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monte Carlo</td>
<td>Sampling</td>
<td>Estimate (8)</td>
</tr>
<tr>
<td>80</td>
<td>$(1.5 \pm 0.3) \times 10^{-2}$</td>
<td>$(1.46 \pm 0.09) \times 10^{-2}$</td>
<td>$1.43 \times 10^{-2}$</td>
</tr>
<tr>
<td>85</td>
<td>$(3 \pm 1) \times 10^{-3}$</td>
<td>$(4.7 \pm 0.3) \times 10^{-3}$</td>
<td>$4.21 \times 10^{-3}$</td>
</tr>
<tr>
<td>90</td>
<td>0</td>
<td>$(1.16 \pm 0.08) \times 10^{-3}$</td>
<td>$1.17 \times 10^{-3}$</td>
</tr>
<tr>
<td>95</td>
<td>0</td>
<td>$(3.2 \pm 0.5) \times 10^{-4}$</td>
<td>$3.11 \times 10^{-4}$</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>$(7.5 \pm 0.5) \times 10^{-5}$</td>
<td>$7.87 \times 10^{-5}$</td>
</tr>
</tbody>
</table>
well with the simulation results and are within two standard errors of the importance sampling estimates.

5 Discussion

We have derived asymptotic approximations for the large exceedance p-values of scan statistics for marked Poisson processes. The proof is based on a change of measure approach, through which we have also designed an importance sampling method for fast Monte Carlo evaluation of p-values. Numerical studies in Section 4 showed that the asymptotic approximation is close to the importance sampling estimates.

We applied our approximations to several classical examples in DNA sequence analysis and found them to be in agreement with previous approximations for the p-values for the unweighted scan statistics. We also calculated the p-values for high scoring windows for a weighted scoring scheme in the search for clusters of palindromes in viral genomes, with each incidence of a palindrome weighted by its length. We assumed a geometric distribution for the palindrome lengths and provided a simple formula for the overshoot constant \( \nu_x \). For more general scoring schemes, we could compute the overshoot constant using (12) or (13) and recursive numerical evaluation. This was illustrated in Example 4 for the shifted geometric distribution.

Our theoretical results depend on the assumption of independence of the \( X_i \)'s. For the problems in genome analysis that we study, this is roughly true because the locations of the events (palindromes in Examples 1 and 2 or DAM sites in Example 3) are spaced on the order of hundreds of bases apart, far enough to escape the documented local dependence in DNA sequences. In support of this assumption, for each of the genomes analyzed we tested the hypothesis that \( S_{N(t)} \) has independent increments using the procedure described in Section
6.2. The p-values for all tests are above the 0.05 level.

A more involved scoring scheme for locating replication origin in viruses could take into account sequence composition, as well as allow (and compensate) for mismatches and gaps in the palindrome. For example, we could count the occurrences of inverted repeats, which are patterns constructed by inserting arbitrary base pairs in the middle of a palindromic pattern. The pattern GGACCTCC is an example of an inverted repeat, constructed by inserting CC in the middle of GGATCC. We then assign scores that reward longer underlying palindromic patterns as well as penalize for the number of inserted base pairs.

Recently there has been increasing interest of scanning entire genomes for regions that are enriched with a certain type of signal, such as transcription factor binding sites (Lifanov et al. (2003), Rajewsky et al. (2002)), segments of high evolutionary conservation (Siepel et al., 2005), or a combination of both (Blanchette et al., 2006). The scoring functions that are used for such scans are often quite complex. For example, in Blanchette et al. (2006), a multiple alignment of the human, mouse and rat genomes is used to locate the “hits”, and the pattern scores of the sequence in all three genomes are combined when computing the score of a hit. Such complex scoring schemes can be analyzed using our methods by choosing the appropriate distribution function $F$.

6 Appendix

6.1 Proof of Theorem 1

Let $0 < m < u$ and define

$$B_{w,m} = \left\{ \sup_{w < t \leq w + m} \left[ S_{N(t+u)} - S_N(t) \right] \geq ux, \left[ S_{N(w+u)} - S_{N(w)} \right] < ux \right\}.$$ (19)
We will show that if \( m \to \infty \) as \( u \to \infty \) with \( m = o(u) \), then for all \( 0 \leq w \leq n - m \),

\[
P(B_{w,m}) = P(B_{0,m}) \sim m \nu x e^{-uI(x)}(x - \lambda \mu) / [2 \pi u \lambda K''(\theta_x)]^{1/2}.
\] (20)

Let \( C > (x - \lambda \mu) \) and \( \ell > 0 \). We will also show that

\[
J_{1,r} := P\{S_{N(rm+u)} - S_{N(rm)} < u \ell - C m, \sup_{rm < t \leq (r+1)m} [S_{N(t+u)} - S_{N(t)}] \geq u \ell \} = o(m e^{-uI(x)} u^{-1/2}),
\]

\( J_{2,r} := P(B_{rm,m} \cap B_{(r+1)m,(\ell-1)m}) = o(m e^{-uI(x)} u^{-1/2}). \) (21)

Moreover, there exists \( \ell > 0 \) (depending on \( C \)) such that for all \( \kappa > 0 \),

\[
J_{3,r} := \sum_{q = \ell}^{\lfloor \kappa u / m \rfloor - 1} P\{S_{N(rm+u)} - S_{N(rm)} \geq u \ell - C m, S_{N(rm+qm+u)} - S_{N(rm+qm)} \geq u \ell - C m\} = o(m e^{-uI(x)} u^{-1/2}),
\] (22)

uniformly over all \( 0 \leq r < \kappa u / m \). We can conclude from the inequalities

\[
\sum_{r=0}^{\lfloor \kappa u / m \rfloor - 1} [P(B_{rm,m}) - J_{1,r} - J_{2,r} - J_{3,r}] \leq P(B_{0,\kappa u}) \leq \sum_{r=0}^{\lfloor \kappa u / m \rfloor} P(B_{rm,m}),
\]

and (20)-(22), that for any fixed \( \kappa > 0 \),

\[
P(B_{0,\kappa u}) \sim \kappa \nu x e^{-uI(x)}(x - \lambda \mu) / [2 \pi u \lambda K''(\theta_x)]^{1/2}.
\] (23)

Theorem 1 then follows from (23) with \( \kappa \) large, the bound

\[
P\{S_{N(t+u)} - S_{N(t)} \geq u x\} = P\{S_{N(u)} \geq u x\} = O(u^{-1/2} e^{-uI(x)}),
\] (24)

and the independence of \( B_{t,u} \) and \( B_{w,\kappa u} \) for \( (w - t) > (\kappa + 1)u \). \( \square \)

Proof of (20). Let us first assume that \( F \) is arithmetic with span 1. Let \( \theta_x > 0 \) and \( \alpha_x > \lambda \) satisfy (3). We shall introduce here the probability measure \( Q \) under which \( N \) is non-uniform Poisson with rate \( \alpha_x \) on \( (0, u] \) and rate \( \lambda \) on \( (u, n] \), and \( X_1, \ldots, X_{N(n)} \) are independent
random variables with $X_j$ having distribution $F_{\theta_x}$ for $1 \leq j \leq N(u)$ and distribution $F$ for $N(u) < j \leq N(n)$. Let $\mathbf{X} = \{N, X_1, \ldots, X_{N(n)}\}$. By (3)-(5),

$$\frac{dQ}{dP}(\mathbf{X}) = \left\{ e^{\theta_x S_N(u)}/[K(\theta_x)]^{N(u)} \right\} e^{-u\alpha_x u}(\alpha_x u)^{N(u)}/[e^{-\lambda u}u N(u)]$$

$$= e^\theta_x S_N(u) - (\alpha_x - \lambda)u = e^{\alpha I(x) + \theta_x (S_N(u) - u)}.$$  \quad (25)

By stationarity,

$$P(B_{u,m}) = P(B_{0,m}) = E_Q[e^{-u \alpha I(x) - \theta_x (S_N(u) - ux)} 1_{B_{0,m}}]$$

$$= e^{-u \alpha I(x)} \sum_{b=1}^{\infty} e^{\theta_x b} Q\left\{ \sup_{0<t\leq m} [S_N(t+u) - S_N(t)] \geq ux | S_N(u) = ux - b \right\}$$

$$\times Q\{S_N(u) = ux - b\},$$ \quad (26)

where the notation $E_Q$ denotes expectation under probability measure $Q$ and $Q(A)$ denotes the probability that the event $A$ occurs under $Q$. It follows from (3) that under $Q$, the sum $S_N(u)$ is asymptotically normal with mean $u\alpha_x K'(\theta_x)/K(\theta_x) = ux$ and variance $u\alpha_x K''(\theta_x)/K(\theta_x) = u\lambda K''(\theta_x)$. Hence for any $b$ fixed,

$$Q\{S_N(u) = ux - b\} \sim [2\pi u \lambda K''(\theta_x)]^{-1/2} \quad \text{as } u \to \infty.$$ \quad (27)

We shall now evaluate the conditional term in (26). Observe that $[S_N(t+u) - S_N(t)]$ is the sum of all $X_j$ lying inside the window $(t, t+u]$ [in other words for all $N(t) < j \leq N(t+u)$]. We slide the window $(t, t+u]$ from $t = 0$ to $t = m$. At each point $t$ whereby there is an inclusion or exclusion of an $X_j$ from the interval $(t, t+u]$, there is a jump in the sum $[S_N(t+u) - S_N(t)]$. We represent the amount of this jump (possibly negative) by some $Y_k$. Conditioned on $S_N(u) = ux - b$, the crossing $[S_N(t+u) - S_N(t)] \geq ux$ occurs if there is an accumulation of jumps exceeding $b$. Since $m = o(u)$ and the Poisson process on $(u, m+u]$ has rate $\lambda$ with observations $X_j$ lying inside $(u, m+u]$ following distribution $F$ while the Poisson process on $(0, m]$ is of rate $\alpha_x$ with observations $X_j$ lying inside $(0, m]$ of distribution $F_{\theta_x}$,

$$Q\left\{ \sup_{0<t\leq m} [S_N(t+u) - S_N(t)] \geq ux \right| S_N(u) = ux - b \right\} \sim P\left\{ \sup_{1\leq k\leq T} R_k \geq b \right\},$$ \quad (28)
where \( T \) is Poisson with mean \( m(\lambda + \alpha_x) \) and \( R_k = Y_1 + \cdots + Y_k \) with \( Y_1, Y_2, \ldots \) i.i.d. random variables such that

\[
P\{Y_1 = y\} = \left(\frac{\lambda}{\lambda + \alpha_x}\right)f(y) + \left(\frac{\alpha_x}{\lambda + \alpha_x}\right)f_{\theta_x}(-y).
\]

Let \( P_* \) be another probability measure under which \( Y_1, Y_2, \ldots \) are i.i.d. with probability mass function \( g \) [see (6)]. Then by (3), (5) and (29),

\[
P_*\{Y_1 = y\} = g(y) = \left(\frac{\alpha_x}{\lambda + \alpha_x}\right)f_{\theta_x}(y) + \left(\frac{\lambda}{\lambda + \alpha_x}\right)f(-y)
= \left(\frac{\alpha_x}{\lambda + \alpha_x}\right)e^{\theta_x y}(\lambda/\alpha_x)f(y) + \left(\frac{\lambda}{\lambda + \alpha_x}\right)e^{\theta_x y}(\alpha_x/\lambda)f_{\theta_x}(-y)
= e^{\theta_x y}P\{Y_1 = y\}. \tag{30}
\]

Let \( E_* \) denote expectation with respect to \( P_* \) and \( \tau_b = \inf\{k : R_k \geq b\} \). Then by (7) and (30),

\[
P\left\{ \sup_{1 \leq k \leq T} R_k \geq b \right\} = E_*\left( e^{-\theta_x R_{\tau_b}} \mathbf{1}_{\{\sup_{1 \leq k \leq T} R_k \geq b\}} \right) \sim \nu_x e^{-\theta_x b} P\left\{ \sup_{1 \leq k \leq T} R_k \geq b \right\}. \tag{31}
\]

By substituting (31) into (28) and then substituting (27), (28) into (26), we obtain

\[
e^{u(x)}[2\pi u \lambda K''(\theta_x)]^{1/2} P(B_{w,m}) \sim \nu_x \sum_{b=1}^{\infty} P_*\left\{ \sup_{1 \leq k \leq T} R_k \geq b \right\} = \nu_x E_*\left( \sup_{1 \leq k \leq T} R_k \right)
\sim \nu_x (E_* R_T) = \nu_x m(\alpha_x E_{\theta_x} X_1 - \lambda \mu), \tag{32}
\]

and (20) follows from (3) because \( E_{\theta_x} X_1 = K'(\theta_x)/K(\theta_x) = (x/\alpha_x) \).

The proof above can also be modified to show (20) when \( F \) is arithmetic with span \( \eta \) by replacing the sum \( \sum_{b=1}^{\infty} \) in (26) by \( \sum_{b \in \mathbb{Z}^+} \). Similarly, when \( F \) has a density function, (20) can be shown by the steps above by changing the sum in (26) by a corresponding integral.

**Proof of (21).** Let \( x^+ = \max(x, 0) \). By substituting (31) into (28) and then substituting (27), (28) into (26), we obtain

\[
e^{u(x)}[2\pi u \lambda K''(\theta_x)]^{1/2} P\left\{ S_{N(m)} < ux - Cm, \sup_{rm < t \leq (r+1)m} S_{N(t)} \geq ux \right\}
\sim \nu_x \sum_{b > Cm} P_*\left\{ \sup_{1 \leq k \leq T} R_k \geq b \right\} \sim m\nu_x E_*\left[ (m^{-1} \sup_{1 \leq k \leq T} R_k - C)^+ \right] = o(m),
\]

17
because $m^{-1} \sup_{1 \leq k \leq T} R_k$ has asymptotic mean $(x - \lambda \mu)(< C)$ and variance converging to zero as $m \to \infty$. Hence the first part of (21) holds.

The second part of (21) follows from (20), (24) and the relation

$$J_{2,r} = P(B_{rm,m} \cup B_{(r+1)m,((r-1)m)} - P(B_{rm,m}) - P(B_{(r+1)m,((r-1)m)}. \quad \text{(33)}$$

\[ \Box \]

**Proof of (22).** By stationarity, $J_{3,r}$ is monotone decreasing with respect to $r$ and hence we need only consider $r = 0$. By the independence of $S_{N(u)}$ and $(S_{N(u+qm)} - S_{N(qm)})$ for $q \geq u/m$,

$$\sum_{q=\lfloor u/m \rfloor + 1}^{\lfloor xu/m \rfloor} P\{S_{N(u)} \geq u x - C m, S_{N(u+qm)} - S_{N(qm)} \geq u x - C m\} = (\lfloor ku/m \rfloor - \lfloor u/m \rfloor)P^2\{S_{N(u)} \geq u x - C m\}. \quad \text{(34)}$$

Let us next consider a fixed $q < u/m$. Let $\tilde{Q}$ be the probability measure under which $X = \{N, X_1, \ldots, X_{N(u)}\}$ is a marked point process with rate $\alpha_x = \lambda \kappa(\theta_x)$ on $(qm, u]$, rate $\tilde{\alpha}_x := \lambda \kappa(\theta_x/2)$ on $[0, qm] \cup (u, qm + u]$ and rate $\lambda$ on $(qm + u, n]$. Moreover, we require that under $\tilde{Q}$,

$$X_j \sim \begin{cases} F_{\theta_x} & \text{for } N(qm) < j \leq N(u) \\ F_{\theta_x/2} & \text{for } j \leq N(qm) \text{ and } N(u) < j \leq N(qm+u) \\ F & \text{otherwise.} \end{cases}$$

By (5),

$$\frac{d\tilde{Q}}{dP}(X|N) = \left\{ e^{\theta_x[S_{N(u)} - S_{N(qm)}]/[K(\theta_x)]^{N(u) - N(qm)}} \right\} \times \left\{ e^{(\theta_x/2)(S_{N(qm)} + S_{N(u+qm)} - S_{N(u)})/[K(\theta_x/2)]^{N(qm) + N(u+qm) - N(u)}} \right\},$$

$$\frac{d\tilde{Q}}{dP}(N) = e^{-2qm(\tilde{\alpha}_x - \lambda)}(\tilde{\alpha}_x/\lambda)^{N(qm) + N(u+qm) - N(u)} \left\{ e^{-(u-qm)(\alpha_x - \lambda)}(\alpha_x/\lambda)^{N(u) - N(qm)} \right\}.$$  

Since $(d\tilde{Q}/dP)(X) = (d\tilde{Q}/dP)(X|N) \times (d\tilde{Q}/dP)(N)$, it follows from (3) that

$$\frac{d\tilde{Q}}{dP}(X) = e^{\theta_x(S_{N(u)} + S_{N(qm+u)} - S_{N(qm)})/2 - 2qm(\tilde{\alpha}_x - \lambda) - (u-qm)(\alpha_x - \lambda)}. \quad \text{(35)}$$
Since $K(\theta_x) \geq [K(\theta_x/2)]^2$, it follows that $(\alpha_x/\lambda) \geq (\tilde{\alpha}_x/\lambda)^2$ and noting that $\tilde{\alpha}_x^2/\lambda - (2\tilde{\alpha}_x - \lambda) = \lambda^{-1}(\tilde{\alpha}_x - \lambda)^2 > 0$, we conclude that

$$(\alpha_x - \lambda) \geq (\tilde{\alpha}_x^2/\lambda - \lambda) > 2(\tilde{\alpha}_x - \lambda). \quad (36)$$

Let $\zeta = (\alpha_x - \lambda) - 2(\tilde{\alpha}_x - \lambda)(> 0)$. By (4), (35) and (36),

$$\frac{d\tilde{Q}}{dP}(X) \geq \exp^{\theta_x(S_{N(u)} + S_{N(qm+u)} - S_{N(qm)})/2 - u(\alpha_x - \lambda) + \zeta q_m} = \exp^{u f(x) + \zeta q_m + \theta_x[(S_{N(u)} + S_{N(qm+u)} - S_{N(qm)})/2 - u x]},$$

and hence by an analogue of (27),

$$P\{S_{N(u)} \geq u x - C m, S_{N(qm+u)} - S_{N(qm)} \geq u x - C m\} \leq P\{S_{N(u)} + S_{N(qm+u)} - S_{N(qm)} \geq 2(u x - C m)\} = E_{\tilde{Q}}\left[\frac{dP}{d\tilde{Q}}(X)1_{\{S_{N(u)} + S_{N(qm+u)} - S_{N(qm)} \geq 2(u x - C m)\}}\right] = O(u^{-1/2}e^{-uf(x) + \theta_x C m - \zeta q_m}), \quad (37)$$

and (22) follows from (24), (34) and (37) by choosing $\ell > \theta_x C/\zeta$. □

### 6.2 A Test for Independence

To test that $S_{N(t)}$ has independent increments, we employed the following procedure:

1. Divide $[0, n]$ into $l$ segments, each of length $n/l$. Set $Z_k = S_{N(kn/l)} - S_{N((k-1)n/l)}$ for $k = 1, \ldots, l$.

2. Divide the state space of $Z_k$ evenly into $r$ blocks $\{R_1, \ldots, R_r\}$. Perform the non-parametric chi-square goodness of fit test for independence of the pairs $(Z_k, Z_{k+1})$ using a $r \times r$ contingency table with entries

$$N_{i,j} = \#\{k : Z_k \in R_i, Z_{k+1} \in R_j\}.$$
Table 6. Test of independent increments for three virus genomes.

<table>
<thead>
<tr>
<th>virus name</th>
<th>n</th>
<th>u</th>
<th>l</th>
<th>r</th>
<th>$\chi^2$</th>
<th>$p-value$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CeHV1</td>
<td>156</td>
<td>789</td>
<td>800</td>
<td>160</td>
<td>198.3317</td>
<td>0.08</td>
</tr>
<tr>
<td>BoHV1</td>
<td>135</td>
<td>301</td>
<td>700</td>
<td>140</td>
<td>48.2131</td>
<td>0.80</td>
</tr>
<tr>
<td>BoHV5</td>
<td>138</td>
<td>390</td>
<td>700</td>
<td>140</td>
<td>11.1809</td>
<td>0.44</td>
</tr>
</tbody>
</table>

The parameter $l$ in this procedure should be chosen to be smaller than the scanning window size $u$, since dependence within this range would cause a “clumping” effect that renders our approximations inaccurate. We set $l$ to be $u/5$. The number of blocks $r$ is set to be $\text{Range}(Z_k : 1 \leq k \leq l - 1)/4$. Table 6 shows the parameters, chi-square statistics and p-values for the three genomes in Example 2.

7 Acknowledgements

We would like to thank an associate editor and two referees for their useful comments, suggestions and references, which have led to a substantial improvement in the presentation of the paper.

References


