Poisson Approximation for Two Scan Statistics with Rates of Convergence

Xiao Fang

(Joint work with David Siegmund)

National University of Singapore

May 28, 2015

Outline

- The first scan statistic
- The second scan statistic
- Other scan statistics

A statistical testing problem

Let $\{X_1, \ldots, X_n\}$ be an independent sequence of random variables. We want to test the hypothesis

$$H_0 : X_1, \ldots, X_n \sim F_{\theta_0}(\cdot)$$

against the alternative

$$H_1 : \text{for some } i < j, X_{i+1}, \ldots, X_j \sim F_{\theta_1}(\cdot)$$
$$X_1, \ldots, X_i, X_{j+1}, \ldots, X_n \sim F_{\theta_0}(\cdot)$$

- i and j are called change-points. They are not specified in the alternative hypothesis.
- θ_0 may be given, or may need to be estimated.
- θ_1 may be given, or may be a nuisance parameter.

The first scan statistic

- If $j - i = t$ is given and $F_{\theta_0}(\cdot)$ and $F_{\theta_1}(\cdot)$ have different mean values, a natural statistic is

$$M_{n;t} = \max_{1 \leqslant i \leqslant n-t-1} T_i, \quad T_i = X_i + \cdots + X_{i+t-1}.$$

- We are interested in its p-value: Assume $X_1, \ldots, X_n \sim F_{\theta_0}(\cdot)$,

$$P(M_{n;t} \geqslant b) = P(\max_{1 \leqslant i \leqslant n-t+1} T_i \geqslant b)$$

$$=?$$

Known results

- Let $Y_i = I(T_i \geqslant b)$.
- $\{\max_{1 \leqslant i \leqslant n-t+1} T_i \geqslant b\} = \{\sum_{i=1}^{n-t+1} Y_i \geqslant 1\}$.
- Dembo and Karlin (1992) proved that if t is fixed and $b, n \to \infty$ plus mild conditions on $F_{\theta_0}(\cdot)$, then

$$P(M_{n;t} \geqslant b) = P(\sum_{i=1}^{n-t+1} Y_i \geqslant 1) \to 1 - e^{-\lambda}$$

  where $\lambda = (n-t+1)E(Y_1)$.
- Mild conditions on $F_{\theta_0}(\cdot)$ ensures that

$$P(Y_{i+1} = 1 | Y_i = 1) \to 0.$$

$t \to \infty$:

- If $X_i \sim$ Bernoulli(p) and b is an integer, Arratia, Gordon and Waterman (1990) prove that

$$|P(M_{n;t} \geqslant b) - (1 - e^{-\lambda})| \leqslant C(e^{-ct} + \frac{t}{n})(\lambda \wedge 1) \quad (1)$$

  where $\lambda = (n - t + 1)P(T_1 = b)(\frac{b}{t} - p)$.

- Haiman (2007) derived more accurate approximations using the distribution function of

$$Z_k := \max\{T_1, \ldots, T_{kt+1}\} \text{ for } k = 1 \text{ and } 2.$$

  The distribution functions of Z_k for $k = 1$ and 2 are only known for Bernoulli and Poisson random variables.

- Our objective is to extend (1) to other random variables.

Preparation for the main result:

- Let $\mu_0 = E(X_1)$. We assume $b = at$ where $a > \mu_0$.

$$P(\max_{1 \leqslant i \leqslant n-t+1} T_i \geqslant b) = P(\max_{1 \leqslant i \leqslant n-t+1} \frac{X_i + \cdots + X_{i+t-1}}{t} \geqslant a).$$

- We assume the distribution of X_1 can be imbedded in an exponential family of distributions

$$dF_\theta(x) = e^{\theta x - \Psi(\theta)} dF(x), \quad \theta \in \Theta. \tag{2}$$

It is known that F_θ has mean $\Psi'(\theta)$ and variance $\Psi''(\theta)$. Assume $\theta_0 = 0$, i.e., $X_1 \sim F$ and there exists $\theta_a \in \Theta^o$ such that $\Psi'(\theta_a) = a$.

- Example: $X_1 \sim N(0,1)$, $\Psi(\theta) = \frac{\theta^2}{2}, \theta_a = a$, $F_{\theta_a} \sim N(a,1)$.

Assumption (2) is used in two places:

1. To obtain an accurate approximation to the marginal probability $P(T_1 \geqslant at)$ by change of measure.
2. Local limit theorem Diaconis and Freedman (1988):

$$d_{TV}(\mathcal{L}(X_1, \ldots, X_m | T_1 = at), \mathcal{L}(X_1^a, \ldots, X_m^a)) \leqslant \frac{Cm}{t}$$

where $X_1^a, \ldots, X_m^a$ are i.i.d. and $X_1^a \sim F_{\theta_a}$.

Let $D_k = \sum_{i=1}^{k}(X_i^a - X_i)$. Let $\sigma_a^2 = \Psi''(\theta_a)$.

Theorem

Under the assumption (2), for some constant C depending only on the exponential family (2), μ_0, and a, we have

$$\left|P(M_{n;t} \geqslant at) - (1 - e^{-\lambda})\right| \leqslant C\left(\frac{(\log t)^2}{t} + \frac{(\log t \wedge \log(n-t))}{n-t}\right)(\lambda \wedge 1),$$

where if X_1 is nonlattice plus mild conditions,

$$\lambda = \frac{(n-t+1)e^{-[a\theta_a - \Psi(\theta_a)]t}}{\theta_a \sigma_a (2\pi t)^{1/2}} \exp[-\sum_{k=1}^{\infty} \frac{1}{k} E(e^{-\theta_a D_k^+})],$$

and if X_1 is integer-valued with span 1,

$$\lambda = \frac{(n-t+1)e^{-(a\theta_a - \Psi(\theta_a))t} e^{-\theta_a(\lceil at \rceil - at)}}{(1-e^{-\theta_a})\sigma_a (2\pi t)^{1/2}} \exp[-\sum_{k=1}^{\infty} \frac{1}{k} E(e^{-\theta_a D_k^+})].$$

Remarks:

- We don't have an explicit expression for the constant C.
- The relative error $\to 0$ if $t, n-t \to \infty$.
- Let $g(x) = Ee^{ixD_1}$ and $\xi(x) = \log\{1/[1-g(x)]\}$.
  Woodroofe (1979) proved that for the nonlattice case,

$$\sum_{k=1}^{\infty} \frac{1}{k} E(e^{-\theta_a D_k^+}) = -\log[(a-\mu_0)\theta_a] - \frac{1}{\pi}\int_0^{\infty} \frac{\theta_a^2[I\xi(x) - \frac{\pi}{2}]}{x(\theta_a^2 + x^2)} dx + \frac{1}{\pi}\int_0^{\infty} \frac{\theta_a\{R\xi(x) + \log[(a-\mu_0)x]\}}{\theta_a^2 + x^2} dx$$

  where R and I denote real and imaginary parts.
  Tu and Siegmund (1999) proved that for the arithmetic case,

$$\sum_{k=1}^{\infty} \frac{1}{k} E(e^{-\theta_a D_k^+}) = -\log(a-\mu_0) + \frac{1}{2\pi}\int_0^{2\pi} \left\{ \frac{\xi(x)e^{-\theta_a - ix}}{1 - e^{-\theta_a - ix}} + \frac{\xi(x) + \log[(a-\mu_0)(1-e^{ix})]}{1-e^{ix}} \right\} dx.$$

Example 1: Normal distribution.

n	t	a	p_1	p_2
1000	50	0.2	0.9315	0.9594
1000	50	0.4	0.2429	0.2624
1000	50	0.5	0.0331	0.0334
2000	50	0.5	0.0668	0.0672

Example 2: Bernoulli distribution.

n	t	μ_0	a	p_1	p_2
7680	30	0.1	$11/30$	0.14097	0.14021
7680	30	0.1	0.4	0.029614	0.029387
15360	30	0.1	0.4	0.058458	0.058003

Sketch of proof:

- Let $m = \lfloor C(\log t \wedge \log(n-t)) \rfloor$. Let

$$Y_i = I(T_i \geqslant at, T_{i+1} < T_i, \ldots, T_{i+m} < T_i \\ T_{i-1} < T_i, \ldots, T_{i-m} < T_i).$$

  Let

$$W = \sum_{i=1}^{n-t+1} Y_i, \quad \lambda_1 = EW = (n-t+1)EY_1.$$

- $P(M_{n;t} \geqslant at) \approx P(W \geqslant 1)$.
- From the Poisson approximation theorem of Arratia, Goldstein and Gordon (1990), we have

$$|P(W \geqslant 1) - (1 - e^{-\lambda_1})| \leqslant C(\frac{1}{t} + \frac{1}{n-t})(\lambda \wedge 1).$$

Approximating λ_1 by λ:

$$\begin{aligned} EY_1 =& P(T_1 \geqslant at, T_2 < T_1, \dots, T_{1+m} < T_1; T_0 < T_1, \dots, T_{1-m} < T_1) \\ \approx& P(T_1 \geqslant at) P^2(T_1 - T_2 > 0, \dots, T_1 - T_{1+m} > 0 | T_1 \approx at) \end{aligned}$$

Note that $T_1 - T_2 = X_1 - X_{t+1}$ and that given $T_1 \approx at$, $X_1 \sim F_{\theta_a}$ approximately and $X_{t+1} \sim F$. Thus,

$$\{T_1 - T_2 > 0\} \approx \{D_1 > 0\} \text{ where } D_1 = X_1^a - X_1.$$

Similarly, $\{T_1 - T_{k+1} > 0\} \approx \{D_k > 0\}$, $D_k = \sum_{i=1}^{k} (X_i^a - X_i)$. Therefore,

$$EY_1 \approx P(T_1 \geqslant at) P^2(D_k > 0, k = 1, 2, \dots).$$

Recall

$$\lambda = \frac{(n-t+1)e^{-[a\theta_a - \Psi(\theta_a)]t}}{\theta_a \sigma_a (2\pi t)^{1/2}} \exp[-\sum_{k=1}^{\infty} \frac{1}{k} E(e^{-\theta_a D_k^+})]. \quad \square$$

Corollary

Let $\{X_1, \ldots, X_n\}$ be i.i.d. random variables with distribution function F that can be imbedded in an exponential family, as in (2). Let $EX_1 = \mu_0$. Assume X_1 is integer-valued with span 1. Suppose $a = \sup\{x : p_x := P(X_1 = x) > 0\}$ is finite. Let $b = at$. Then we have, with constants C and c depending only on p_a,

$$\left|P(M_{n;t} \geqslant b) - (1 - e^{-\lambda})\right| \leqslant C(\lambda \wedge 1)e^{-ct}$$

where

$$\lambda = (n - t)p_a^t(1 - p_a) + p_a^t.$$

The second scan statistic

Recall that we want to test

$$H_0 : X_1, \ldots, X_n \sim F_{\theta_0}(\cdot)$$

against the alternative

$$H_1 : \text{for some } i < j, X_{i+1}, \ldots, X_j \sim F_{\theta_1}(\cdot)$$
$$X_1, \ldots, X_i, X_{j+1}, \ldots, X_n \sim F_{\theta_0}(\cdot)$$

Now assume $j - i$ is not given, and F_{θ_0} and F_{θ_1} are from the same exponential family of distributions

$$dF_\theta(x) = e^{\theta x - \Psi(\theta)} dF(x), \quad \theta \in \Theta.$$

Then the log likelihood ratio statistic is

$$\max_{0 \leqslant i < j \leqslant n} \sum_{k=i+1}^{j} (\theta_1 - \theta_0)(X_k - \frac{\Psi(\theta_1) - \Psi(\theta_0)}{\theta_1 - \theta_0}).$$

It reduces to the following problem:
Let $\{X_1, \ldots, X_n\}$ be independent, identically distributed random variables. Let $EX_1 = \mu_0 < 0$. Let $S_0 = 0$ and $S_i = \sum_{j=1}^{i} X_j$ for $1 \leqslant i \leqslant n$. We are interested in the distribution of

$$M_n := \max_{0 \leqslant i < j \leqslant n} (S_j - S_i).$$

Iglehart (1972) observed that it can be interpreted as the maximum waiting time of the first n customers in a single server queue.

Karlin, Dembo and Kawabata (1990) discussed genomic applications.

The limiting distribution was derived by Iglehart (1972):
Assume the distribution of X_1 can be imbedded in an exponential family of distributions

$$dF_\theta(x) = e^{\theta x - \Psi(\theta)} dF(x), \quad \theta \in \Theta.$$

Assume $EX_1 = \Psi'(0) = \mu_0 < 0$ and there exists a positive $\theta_1 \in \Theta$ such that

$$\Psi'(\theta_1) = \mu_1, \quad \Psi(\theta_1) = 0.$$

When X_1 is nonlattice, we have

$$\lim_{n\to\infty} P(M_n \geqslant \frac{\log n}{\theta_1} + x) = 1 - \exp(-K^* e^{-\theta_1 x}).$$

Theorem

Let $h(b) > 0$ be any function such that $h(b) \to \infty$, $h(b) = O(b^{1/2})$ as $b \to \infty$. Suppose $n - b/\mu_1 > b^{1/2}h(b)$. We have,

$$\left|P(M_n \geqslant b) - (1 - e^{-\lambda})\right| \leqslant C\lambda \left\{ \left(1 + \frac{b/h^2(b)}{n - b/\mu_1}\right) e^{-ch^2(b)} + \frac{b^{1/2}h(b)}{n - \frac{b}{\mu_1}} \right\}$$

where if X_1 is nonlattice plus mild conditions,

$$\lambda = (n - \frac{b}{\mu_1}) \frac{e^{-\theta_1 b}}{\theta_1 \mu_1} \exp(-2 \sum_{k=1}^{\infty} \frac{1}{k} E_{\theta_1} e^{-\theta_1 S_k^+}),$$

and if X_1 is integer-valued with span 1 and b is an integer,

$$\lambda = (n - \frac{b}{\mu_1}) \frac{e^{-\theta_1 b}}{(1 - e^{-\theta_1})\mu_1} \exp(-2 \sum_{k=1}^{\infty} \frac{1}{k} E_{\theta_1} e^{-\theta_1 S_k^+}).$$

Remarks:

- By choosing $h(b) = b^{1/2}$, we get

$$|P(M_n \geqslant b) - (1 - e^{-\lambda})| \leqslant C\lambda\{e^{-cb} + \frac{b}{n}\}$$

- By choosing $h(b) = C(\log b)^{1/2}$ with large enough C, we can see that the relative error in the Poisson approximation goes to zero under the conditions

$$b \to \infty, \quad (b \log b)^{1/2} \ll n - b/\mu_1 = O(e^{\theta_1 b}),$$

  where $n - b/\mu_1 = O(e^{\theta_1 b})$ ensures that λ is bounded.
- For the smaller range (in which case $\lambda \to 0$)

$$b \to \infty, \quad \delta b \leqslant n - b/\mu_1 = o(e^{\frac{1}{2}\theta_1 b})$$

  for some $\delta > 0$, Siegmund (1988) obtained more accurate estimates by a technique different from ours.

Let $G(z) = \sum_0^\infty p_k z^k + \sum_1^\infty q_k z^{-k}$, and let z_0 denote the unique root > 1 of $G(z) = 1$. For the case $p_k = 0$ for $k > 1$, using the notation $Q(z) = \sum_k q_k z^k$, one can show for large values of n and b that $\lambda \sim n z_0^{-b}\{[Q(1) - Q(z_0^{-1})] - (1 - z_0^{-1}) z_0^{-1} Q'(z_0^{-1})\}$. For the case $q_k = 0$ for $k > 1$, $\lambda \sim n z_0^{-b}(1 - z_0^{-1}))|G'(1)|^2/G'(z_0)$. In particular if $q_1 = q$ and $p_1 = p$, where $p + q = 1$, both these results specialize to $\lambda \sim n(p/q)^b(q - p)^2/q$.

Sketch of proof (for the case $h(b) = b^{1/2}$):

- Recall $S_i = \sum_{k=1}^{i} X_k$. Define $T_b := \inf\{n \geqslant 1 : S_n \notin [0, b)\}$.
- For a positive integer m, let ω_m^+ be the m-shifted sample path of $\omega := \{X_1, \dots, X_n\}$. Let $t = \lceil \frac{b}{\mu_1} + b \rceil$ and $m = \lfloor cb \rfloor$ such that $m < t$.
- For $1 \leqslant i \leqslant n - t$, let

$$Y_i = \mathrm{I}\big(S_i < S_{i-j}, \forall\ 1 \leqslant j \leqslant m;\ T_b(\omega_i^+) \leqslant t,\ S_{T_b}(\omega_i^+) \geqslant b\big).$$

  That is, Y_i is the indicator of the event that the sequence $\{S_1, \dots S_n\}$ reaches a local minimum at i and the i-shifted sequence $\{S_i(\omega_\alpha^+)\}$ exits the interval $[0, b)$ within time t and the first exiting position is b.
- Let $W = \sum_{i=1}^{n-t} Y_i$.

Sketch of proof (cont.)

- $P(M_n \geqslant b) \approx P(W \geqslant 1)$.
- $|P(W \geqslant 1) - (1 - e^{-\lambda_1})| \leqslant C\lambda e^{-cb}$.
- $\lambda_1 = (n-t)EY_1 \approx (n-t)P(\tau_0 = \infty)P(S_{T_b} \geqslant b)$ where $\tau_0 := \inf\{n \geqslant 1 : S_n \geqslant 0\}$.
- $\lambda_1 \approx \lambda$.

Other statistics

Recall again that we want to test

$$H_0 : X_1, \ldots, X_n \sim F_{\theta_0}(\cdot)$$

against the alternative

$$H_1 : \text{for some } i < j, X_{i+1}, \ldots, X_j \sim F_{\theta_1}(\cdot)$$
$$X_1, \ldots, X_i, X_{j+1}, \ldots, X_n \sim F_{\theta_0}(\cdot)$$

1. If θ_0 is not given, we need to consider

$$P(M_{n;t} \geqslant b | S_n) \text{ and } P(M_n \geqslant b | S_n).$$

2. If θ_0 is given but θ_1 is a nuisance parameter, then the log likelihood ratio statistic is

$$\max_{0 \leqslant i < j \leqslant n} \max_{\theta} [\theta(S_j - S_i) - (j - i)\Psi(\theta)].$$

For normal distribution, it reduces to

$$\max_{0 \leqslant i < j \leqslant n} \frac{(S_j - S_i)^2}{2(j - i)}.$$

The limit of is only know for normal distribution and for $n \asymp b^2$ [Siegmund and Venkatraman (1995)].

3. Frick, Munk and Sieling (2014) proposed the following multiscale statistic:

$$\max_{0 \leqslant i < j \leqslant n} \left\{ \frac{|S_j - S_i|}{\sqrt{j-i}} - \sqrt{2 \log\left(\frac{n}{j-i}\right)} \right\}.$$

The penalty term $\sqrt{2\log(n/(j-i))}$ was first studied in Dümbgen and Spokoiny (2001) and motivated by Lévy's modulus of continuity theorem.

4. Let $X_1, \dots, X_m$ be an independent sequence of Gaussian random variables with mean $EX_i = \mu_i$ and variance 1. We are interested in testing the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_m$$

against the alternative hypothesis that there exist $1 \leqslant \tau_1 < \cdots < \tau_K \leqslant m-1$ such that

$$\begin{aligned} H_1 : & \mu_1 = \dots \mu_{\tau_1} \neq \mu_{\tau_1+1} = \cdots = \mu_{\tau_2} \neq \cdots = \mu_{\tau_K} \\ & \neq \mu_{\tau_K+1} = \cdots = \mu_m. \end{aligned}$$

4. (cont.)

- If $K = 1$, the log likelihood ratio statistic is

$$\max_{1 \leqslant t \leqslant m-1} \frac{\left| \frac{S_m - S_t}{m-t} - \frac{S_t}{t} \right|}{\sqrt{\frac{1}{t} + \frac{1}{m-t}}}.$$

- If $K > 1$, an appropriate statistic is

$$\max_{0 \leqslant i < j < k \leqslant m} \left\{ \frac{\left| \frac{S_j - S_i}{j-i} - \frac{S_k - S_j}{k-j} \right|}{\sqrt{\frac{1}{j-i} + \frac{1}{k-j}}} \right\}.$$

Thank you!