Journal of Statistical Planning and Inference 190 (2017) 52–59

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Approximations of distributions of scan statistics of inhomogeneous Poisson processes

Tung-Lung Wu

Department of Mathematics and Statistics, Mississippi State University, Starkville, MS 39759, USA

ARTICLE INFO

Article history:
Received 29 August 2016
Received in revised form 23 January 2017
Accepted 29 April 2017
Available online 19 May 2017

MSC:
60E05
60J10

Keywords:
Weighted scan statistics
Finite Markov chain imbedding
Inhomogeneous Poisson process
Discrete approximation

ABSTRACT

In this paper, the distributions of scan statistics of inhomogeneous Poisson processes are studied. First, the distribution of the continuous scan statistic of an inhomogeneous Poisson process is approximated by the distribution of the discrete scan statistic for a sequence of Bernoulli trials with unequal probabilities of success. Next, we introduce a weighted scan statistic to adjust the inhomogeneity and the weighted continuous scan statistic is also approximated by the weighted discrete scan statistic. The finite Markov chain imbedding technique is used to obtain the exact distributions of weighted discrete scan statistics. An example using the weighted scan statistic of an inhomogeneous Poisson process for detecting DNA copy number variation is given. Numerical results and simulations are also given to illustrate our theoretical results.

1. Introduction

Let $N^*(t)$ be a homogeneous Poisson process with intensity λ on $t \in (0, 1]$. For $0 < \omega \leq 1$, let $S^*(\omega, t) = N^*(t+\omega) - N^*(t)$ denote the number of events that have occurred in the interval $(t, t+\omega]$, where ω is the window size. The classic continuous scan statistic is defined as

$$S^*(\omega) = \sup_{0<t\leq 1-\omega} S^*(\omega, t). \tag{1}$$

Various bounds and approximations have been derived for the distribution of $S^*(\omega)$. For example, the upper and lower bounds are given by Janson (1984), and Alm (1999) and Haiman (2000) derived accurate approximations. Fu et al. (2012) derived discrete approximations based on the finite Markov chain imbedding (FMCI) technique and provided the rate of convergence. Applications of scan statistics can be found in many areas such as quality control, system reliability, image analysis and sensor network (see, e.g., Cressie, 1991, Hoh and Ott, 2000, Koutras et al., 1993, Rakitzis and Antzoulakos, 2011, Rosenfeld, 1978, Wallenstein and Naus, 1974). For an extensive review of recent advances on scan statistics generated by homogeneous Poisson processes and Bernoulli trials, see the two books by Glaz and Balakrishnan (1999) and Glaz et al. (2009a).

DNA copy number variants (CNVs) can cause cancer or genetic disease (see, e.g., Chen et al., 2010), and it is a predictive biomarker for many diseases, such as Parkinson's disease (see, e.g., Pyle et al., 2016). In a traditional analysis, the cellular DNA is assumed to be homogeneous and hence the occurrence of the biomarkers is modeled by a homogeneous Poisson process (see, e.g., Castle et al., 2010). A recent work by Shen and Zhang (2012) proposed an inhomogeneous Poisson process model for DNA copy numbers. This motivates us to consider an inhomogeneous Poisson process model for scan statistic. Despite

E-mail address: tw1475@msstate.edu.

http://dx.doi.org/10.1016/j.jspi.2017.04.002

the vast literature on scan statistic in the past three decades, very little work has been done for the scan statistic of an inhomogeneous Poisson process. One exception is the work of Karlin and Chen (2004) who derived a Poisson approximation for the distribution of the kth minimal r-scan of an inhomogeneous Poisson process in $(0, t)$ as $t \to \infty$. The scan statistic is considered as a special case of the kth minimal r-scan. However, no numerical results and comparisons were done to justify their result. Also the adjustment to the inhomogeneity has not been addressed. To the best of our knowledge, currently there is no general method to obtain the distributions of scan statistics of inhomogeneous Poisson processes. To fill this gap, we derive the distributions of scan statistics generated by inhomogeneous Poisson processes and Bernoulli trials.

Let $N(t)$ be an inhomogeneous Poisson process with intensity $\lambda(t)$ on $t \in (0, 1]$ and $m(t) = \int_0^t \lambda(s)ds$. Throughout this paper, we assume that $\lambda(t)$ is a continuous function, and hence is bounded on $(0, 1]$. Let $S(\omega, t) = N(t+\omega) - N(t)$ and define the scan statistic of an inhomogeneous Poisson process as

$$S(\omega) = \sup_{0<t\leq 1-\omega} S(\omega, t). \tag{2}$$

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli trials with $P(X_i = 1) = p_i$ and $P(X_i = 0) = q_i = 1 - p_i$, $i = 1, \ldots, n$. For $1 \leq r \leq n$, the discrete version of the scan statistic can be similarly defined as

$$S_n(r) = \max_{1\leq t\leq n-r+1} S_n(r, t), \tag{3}$$

where $S_n(r, t) = \sum_{k=t}^{t+r-1} X_k$. The first part of this paper is to provide a discrete approximation that for $0 < \omega \leq 1$,

$$P(S(\omega) > a) = \lim_{n\to\infty} P(S_n([n\omega] + k) > a), \tag{4}$$

for any fixed integer k, where $[n\omega]$ is the integer part of $n\omega$ and the relationship between p_i and $\lambda(t)$ is given in Section 2. If the underlying process is not homogeneous, then the classic scan statistic is no longer appropriate. It needs to be adjusted to account for the inhomogeneity. In the second part of this paper, we introduce the weighted scan statistic defined as follows:

$$SW(\omega) = \sup_{0<t\leq 1-\omega} w(t)S(\omega, t), \tag{5}$$

where $w(t)$ is the weight function. A natural choice of the weight function is $w(t) \propto 1/\lambda(t)$, where $\propto$ stands for "proportional to". When $w(t) = 1$, the weighted scan statistic in (5) reduces to the classic scan statistic of an inhomogeneous Poisson process. Throughout the paper, $w(t)$ is assumed to be bounded.

In Section 2, we derive the discrete approximation for the distributions of scan statistics of inhomogeneous Poisson processes. The generalization of classic scan statistics to weighted scan statistics is given in Section 3. In Section 4, an example is given to illustrate the use of the weighted scan statistic of an inhomogeneous Poisson process. Numerical results and simulations are also given in this section. Summary is given in Section 5.

2. Scan statistics of inhomogeneous Poisson processes

Given a large integer n, the interval $(0, 1]$ is divided into n subintervals $(0 = t_0, t_1], \ldots, (t_{n-1}, t_n = 1]$, each of equal length $t_i - t_{i-1} = \Delta t = 1/n$, $i = 1, \ldots, n$. It follows from the definition that the probability of having at least two points in an interval of length Δt is negligible and the probability of having exactly one point is approximately $\lambda(t)\Delta t$. The following lemmas similar to Lemmas 3.1 and 3.2 in Fu et al. (2012) will be used to prove our main Theorem 2.4. For simplicity of notation, with understanding, the last scanning window stops at time $t = 1$.

Lemma 2.1. *Given a window size $0 < \omega \leq 1$, the following holds*

$$\begin{aligned}\sup_{0<t\leq 1-\omega+2\Delta t} S(\omega - 2\Delta t, t) &\leq \max_{1\leq i\leq n-[n\omega]} S_n([n\omega], i) \leq \sup_{0<t\leq 1-\omega} S(\omega, t)\\ &\leq \max_{1\leq i\leq n-[n\omega]-1} S_n([n\omega]+2, i) \leq \sup_{0<t\leq 1-\omega-2\Delta t} S(\omega + 2\Delta t, t).\end{aligned}$$

We omit the proof since it directly follows from the proof of Lemma 2.1 in Fu et al. (2012).

Lemma 2.2. *Given $0 < \omega \leq 1$, $a < N$ and two fixed integers $k > 0$ and $\ell > 0$, we have*

$$\begin{aligned}&|P(S(\omega + k\Delta t, t) > a|N) - P(S(\omega - \ell\Delta t, t) > a|N)| = (a+1)2^N \cdot O\left(\frac{1}{n}\right),\\ &\textit{for } 0 < t < 1 - w + \ell\Delta t.\end{aligned} \tag{6}$$

Proof. Given $N(1) = N$, it follows that, for $t_2 > t_1$, $N(t_2) - N(t_1) \sim B(N, p)$, where $p = \int_{t_1}^{t_2} \lambda(s)ds/m(1)$ and we have

$$\begin{aligned}&|P(S(\omega + k\Delta t, t) > a|N) - P(S(\omega - \ell\Delta t, t) > a|N)|\\ &= P(S(\omega + k\Delta t, t) > a, S(\omega - \ell\Delta t, t) \leq a|N)\end{aligned}$$