Computational Statistics and Data Analysis 82 (2015) 164–172

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A new adaptive procedure for multiple window scan statistics

Tung-Lung Wu *, Joseph Glaz

Department of Mathematics and Statistics, Mississippi State University, USA
Department of Statistics, University of Connecticut, USA

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form 25 August 2014
Accepted 6 September 2014
Available online 16 September 2014

Keywords:
Adaptive procedure
Clusters
Multiple window scan statistics
Sequential
Window size
Two-dimensional

ABSTRACT

Scan statistics have been widely applied to test for unusual cluster of events in many scientific areas. It has been of practical interest on how to select the window size of a scan statistic. An adaptive procedure for multiple window scan statistics is proposed and the distributions are studied for independent identically distributed Bernoulli trials and uniform observations on (0, 1) in one-dimensional case. The idea of the procedure is to select the window sizes sequentially. An initial window size is chosen and the subsequent window sizes are then determined, depending on the value of the current scan statistic at each stage. The power of scan statistics based on the adaptive procedure is compared with power of standard scan statistics. Numerical results and applications for disease clusters detection are given to illustrate our procedure.

1. Introduction

The literature of scan statistics has grown rapidly in the past decade. Many applications involving scan statistics in one or two dimensions have been found in many areas. In one-dimensional case, the scan statistic has been applied by Hoh and Ott (2000) and Lachenbruch et al. (2005) to locate susceptibility genes and to monitor manufactured blood products, respectively. In two-dimensional case, a spatial scan statistic has been studied in detail by Kulldorff (1997) and applied in epidemiology. In systems reliability, Barbour et al. (1996) and Yamamoto and Akiba (2005) studied the reliability of 2-dimensional consecutive k-out-of-n system and 2-dimensional rectangular k-within-consecutive-(r, s)-out-of-(m, n):F system, respectively. When the components can only have two outcomes, success and failure, the reliability of a 2-dimensional system is equivalent to a 2-dimensional scan statistic problem. Other examples are such as image analysis (Rosenfeld, 1978), signal detection in a sensor network (Song et al., 2012) and DNA sequence analysis (Karwe and Naus, 1997). Two books by Glaz et al. (2009) and Glaz and Balakrishnan (1999) provide an excellent overview for recent developments and advances in scan statistics.

The generalized likelihood ratio test will reject the null hypothesis of homogeneity against a cluster alternative if the scan statistic is too large. To apply a fixed window scan statistic, one needs to specify in advance the window size, while the cluster size is usually unknown. Loader (1991) and Nagarwalla (1996) generalized the fixed window scan statistic to a variable window scan statistic where all windows sizes in a given interval are considered. The price for this generalization is that the distribution of a variable window scan statistic can only be obtained using Monte Carlo simulation which is usually time-consuming. In another slightly different direction, Naus and Wallenstein (2004) and Wu et al. (2013) studied multiple window scan statistics, where a fixed number of window sizes are given, and provided the approximate and exact distributions.

In this paper, we propose an adaptive procedure for multiple window scan statistics without specifying the window sizes except for the initial one. Following the procedure, the data-dependent window sizes are determined sequentially according

* Corresponding author at: Department of Mathematics and Statistics, Mississippi State University, USA.
E-mail address: comehome1981@gmail.com (T.-L. Wu).

http://dx.doi.org/10.1016/j.csda.2014.09.002
0167-9473/

to the value of the current fixed window scan statistic at each stage. The procedure continues until a certain criterion is met. To illustrate the statistic of interest, consider a conditional continuous 2-window scan statistic. Let the initial window size be r_1 and the implementation of an adaptive procedure relies on the following probability

$$P(S_N(r_1) \geq s_1 \text{ or } S_N(f(S_N(r_1))) \geq s_2), \tag{1}$$

where $S_N(r)$ is the conditional scan statistic with window size r given the total number of points equal to N, f can be any suitably chosen real-valued function and s_1 and s_2 are chosen to achieve a given significance level. Throughout this paper, the distribution of an adaptive scan statistic is of a similar form in (1).

In Section 2, the general adaptive procedure is described. The approximations of the distributions of the adaptive 2-stage scan statistics for one-dimensional case are also given, including conditional continuous scan statistics and conditional and unconditional discrete scan statistics. The extension to two-dimensional case is given in Section 3. A simple algorithm to find the two-dimensional scan statistic for a given data set is provided. The power comparison between adaptive scan statistics and classic fixed window scan statistics is given in Section 4. In Section 5, applications for disease clusters detection are given to illustrate our procedure for both one and two-dimensional cases. Summary and discussion are given in Section 6.

2. One-dimensional case

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

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

Given the total number of events $N(1) = N$, the dependency on λ is removed. The N arrival times are independent uniformly distributed on $[0, 1)$ and (2) becomes a conditional continuous scan statistic, denoted by $S_N(r)$. As the number of events is usually known, the conditional continuous scan statistic (uniform observations) has been widely studied and therefore is the focus in the continuous cases in this paper. The discrete version of conditional and unconditional scan statistics can be defined similarly.

Given a function f, the adaptive procedure for a one-dimensional multiple window scan statistic is described as follows:

(i) choose an initial window size r_1;
(ii) obtain the value of the scan statistic, say, $S_N(r_1) = s_1$;
(iii) choose $r_2 = f(s_1)$ as the next window size;
(iv) obtain the value of the scan statistic, say, $S_N(r_2) = s_2$;
(v) repeat steps (iii) and (iv) until finished.

In this paper, we consider two types of adaptive procedures. The first one is the k-stage adaptive procedure where the procedure will stop until k window sizes have been specified. The distribution of the resulting scan statistic is

$$P(S_N(r_1) \geq s_1 \text{ or } S_N(r_2) \geq s_2 \text{ or } \cdots \text{ or } S_N(r_k) \geq s_k), \tag{3}$$

where $r_j = f(S_N(r_{j-1}))$, $j = 2, \ldots, k$. The special case $k = 2$ is given in (1) and will be studied in detail for both continuous and discrete cases. The second adaptive procedure is to repeat the step (v) until a certain criterion is satisfied. For example, a p-value is calculated for the fixed window scan statistic at each stage. We continue the procedure as long as the successive p-values are non-increasing (Hoh and Ott, 2000). In other words, the procedure will stop if the p-value starts increasing. We call this the minimum p-value procedure.

2.1. Continuous case: uniform observations

Given a function f, we derive an approximation for the distribution in (1) when the observations are taken from the uniform distribution on $[0, 1)$. Throughout this section, the specific function $f(S_N(r_1)) = r_1 S_N(r_1)/E(S_N(r_1))$ is discussed, where $E(S_N(r_1))$ is the expected value of $S_N(r_1)$. For simplicity, the second window size $f(S_N(r_1))$ induced by the value of the first scan statistic is denoted by r_2 throughout this paper.

A 2-stage procedure. Given $N(1) = N$, the N points are randomly distributed according to the uniform distribution on $[0, 1)$. An adaptive 2-stage procedure for a continuous scan statistic is described as follows:

(i) choose an initial window size r_1;
(ii) obtain the value of the scan statistic, say, $S_N(r_1) = s_1$;
(iii) choose $r_2 = r_1 s_1/E(S_N(r_1))$ as the next window size;
(iv) obtain the value of the scan statistic, say, $S_N(r_2) = s_2$.

Based on the above procedure, the probability we are interested in is given by

$$P(S_N(r_1) < s_1, S_N(r_2) < s_2) = \sum_{x=0}^{s_1-1} P\left(S_N(r_2) < s_2 | S_N(r_1) = x\right) P(S_N(r_1) = x). \tag{4}$$

Note that r_2 depends on r_1. First we assume $r_2 < r_1$. The probability $P(S_N(r_1) = x)$ has been obtained by many authors either exactly or approximately. For example, the exact probabilities for limited values of r_1 are tabulated in Neff and Naus (1980).