

# Research Statement

Shihua Wen (2007)

## Introduction

Scan statistics arise when scanning in time or space, or both, looking for clusters of certain *events* or *cases*. Here, *event* means the occurrence of some type of disease or crime, or some sort of physical or chemical measurement, etc. A *cluster* can be defined as a certain spatial or temporal subregion where the the probability distribution of an event is different from that in the rest of the region. For instance, a city neighborhood where the crime rate is higher than in the rest of the city defines a cluster. If clusters can be detected more accurately, better decisions or more efficient policies can be made. More examples can also be found in many fields, including epidemiology, criminology, genetics, mining, astronomy, and so on.

The modern literature about scan statistics can be traced back to the 1960's. At present, one of the most popular methods is Kulldorff's scan statistics method which bears the name of its developer (Kulldorff, 1997). Kulldorff's scan statistics method imposes a movable variable-size scan window to detect both the location and the size of the clusters in purely spatial or higher dimensions, e.g. space-time, by adjusting for uneven background population. It requires assumptions on the underlying distribution (Bernoulli, Poisson, or exponential, etc.) of the scanned region. In addition, the method requires knowledge of the number of events over the region of interest for Monte Carlo hypothesis testing.

In my doctoral research, I developed a semi-parametric scan statistics method to detect clusters. This method can be regarded as a certain semi-parametric generalization of Kulldorff's method which requires much less than complete distributional assumptions, and which does not require the number of cases prior to scanning. Both empirical simulation power studies and real data analysis have demonstrated the value of this method.

## Method and Results

The Semi-parametric approach I utilized in my research is based on a density ratio model studied by Fokianos et al. (2001), and Qin and Zhang (1997). This model is ideal for testing equidistribution given two or more samples. Following a similar scanning scheme as Kulldorff's, consider a scanning window which separates the study region into two parts ( $m = 2$ ),

$$\begin{aligned}\mathbf{x}_1 &= (x_{11}, x_{12}, \dots, x_{1n_1})' \sim g_1(x) \\ \mathbf{x}_2 &= (x_{21}, x_{22}, \dots, x_{2n_2})' \sim g_2(x)\end{aligned}$$

where  $\mathbf{x}_1$  is the sample inside the scanning window with sample size  $n_1$ , and the  $\mathbf{x}_2$  is the sample outside the scanning window with sample size  $n_2$ ,  $g_j(x)$  is the probability density function of  $x_{ji}, j = 1, 2; i = 1, \dots, n_j$ . Choosing the sample outside the scanning window as the reference sample and  $g_2(x)$  as the reference density, it is assumed that the density ratio between the density inside the scanning window and the reference density has an exponential form

$$\frac{g_1(x)}{g_2(x)} = \exp\{\alpha_1 + \beta_1' \mathbf{h}(x)\}.$$

Here  $\mathbf{h}(x)$  is a known vector-valued function of  $x$  which may take on a scalar form such as  $x$  or a vector-valued form such as  $(x, x^2)'$ ,  $\alpha_1$  is a scalar, but  $\beta_1$  could be a scalar or vector depending on  $\mathbf{h}(x)$ . Clearly,  $\beta_1 = \mathbf{0}$  implies  $\alpha_1 = 0$ . Therefore, testing the null hypothesis of no cluster means to test  $\beta_1 = \mathbf{0}$ . Three semi-parametric test statistics,  $\chi_1$  test statistic,  $\chi_2$  test statistic, and likelihood ratio test statistic, have been proposed for such hypothesis testing (Kedem and Wen, 2007).

Following a similar scanning procedure as Kulldorff's, the semi-parametric method generates a large set of overlapping scanning windows and performs each window a semi-parametric two-sample test. Since each test corresponds to a test statistic and its  $p$ -value, it induces a multiple-testing problem. I use Storey's  $q$ -value method, which is a type of false discovery rate (FDR) method, to take account of this problem (Benjamini and Hochberg, 1995; Storey, 2002). The clusters are then the scanning windows with significant  $q$ -values, i.e,  $q < 0.05$ .

The semi-parametric approach has several advantages as follows,

- ① The reference (or background) distribution and all the parameters are estimated from the combined data, not just from a single sample either inside the window or outside the window.
- ② For a properly chosen  $\mathbf{h}(x)$ , the above tests are more powerful than  $t$ -test or  $F$ -test (Gagnon, 2005; Fokianos et al., 2001). Under certain conditions, the semi-parametric scan statistics method is more powerful than kulldorff's method.
- ③ In testing equidistribution, other than an assumption regarding the tilt function  $\mathbf{h}(t)$ , the semi-parametric density ratio method does not require distributional assumptions. It seems that for a non-homogeneous regional variance the choice of  $\mathbf{h}(x) = (x, x^2)'$  suggested by the normal distribution is sensible.
- ④ The semi-parametric method can be applied to either continuous or discrete distributions.
- ⑤ Since the asymptotic distributions of the above mentioned test statistics are known, in principle there is no need for the time consuming Monte-Carlo methods to compute the  $p$ -values.

It is shown by empirical simulation power studies that when data are binary or satisfied with Poisson assumption, the semi-parametric method achieves good power comparable to that achieved by Kulldorff's method and by a certain focused test in testing the hypothesis of no cluster (Waller and Lawson, 1995; Kedem and Wen, 2007). When the data are not binary, such as ordinal categorical data, the semi-parametric method still works in many cases, but Kulldorff's method requires the choices of a correct probability model, namely the correct scan statistic, in order to achieve power comparable to that achieved by the semi-parametric method. Kulldorff's method with an inappropriate probability model may lose power. I have applied the semi-parametric method to childhood leukemia and lymphoma data set in North Humberside, England. The semi-parametric method correctly detected the cluster suggested by the medical and epidemiology studies.

### **Future Work**

I am currently using the circular scan window to search clusters. In my future research, I plan to extend the semi-parametric method to scan clusters of other shapes or irregular shape, by adopting other scanning procedures, such as elliptic window scan (Kulldorff et al., 2006), upper level set scan (Patil and Taillie, 2004), or flexible scan (Tango and Takahashi, 2005). Another potential improvement of semi-parametric method is to incorporate covarites, such as demographic information. I believe it will be challenging but also promising.

### **Research Interests**

My research interests lie in a broad field in applied and computational statistics, including the semi-parametric method which I am currently working on, spatial statistics, time series analysis, Bayesian statistics, and data mining techniques and applications.

### **References**

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, Vol. 57, No. 1, 289-300.

Fokianos, K., Kedem, B., Qin, J., and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometric*, 43, 56-65.

Gagnon, R. (2005), Certain Computational Aspects of Power Efficiency and of State

Space Models. *Ph.D Dissertation, Department of Mathematics, University of Maryland, College Park.*

Kedem, B., Wen, S. (2007) Semi-parametric Cluster Detection. *Journal of Statistical Theory and Practice*, Vol. 1, No. 1 (inaugural issue).

Kulldorff, M. (1997). A spatial scan statistic. *Communication in Statistics: Theory and methods*, 26, 1481-1496.

Patil, G.P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11, 183-197.

Qin, J., and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609-618.

Storey J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64: 479-498.

Tango, T. and Takahashi K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4: 11.

Waller, L.A. and Lawson, A.B. (1995). The power of focused tests to detect disease clustering. *Statistics in Medicine*, 14, 2291-2308.