Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Wald-based spatial scan statistics for cluster detection

Ying Liu [a,*], Yawen Liu [a], Tonglin Zhang [b]

[a] School of Statistics, University of International Business and Economics, No. 10, Huixin Dongjie, Chaoyang District, Beijing 100029, China

[b] Department of Statistics, Purdue University, 250 N University St., West Lafayette, IN 47907-2066, USA

ARTICLE INFO

Article history:
Received 22 July 2017
Received in revised form 3 June 2018
Accepted 4 June 2018
Available online xxxx

Keywords:
Cluster detection
Generalized linear models
Likelihood ratio
Overdispersion
Spatial scan statistics
Wald-based statistics

ABSTRACT

The spatial scan test, which is often carried out by maximizing a likelihood ratio-based statistic over a collection of cluster candidates, is widely used in cluster detection and disease surveillance. As the likelihood ratio statistic may not be available if the exact distribution of the response variable is not specified, a Wald-based spatial scan approach is proposed. The idea is to construct a special explanatory variable for spatial clusters in the linear function of a statistical model. The spatial scan test is carried out by scanning the special explanatory variable over the collection of cluster candidates. An advantage is that the Wald-based spatial scan statistic can bridge spatial clusters and linear functions of statistical models. It can be easily combined with well-known statistical models beyond generalized linear models. It is expected that the proposed approach will have a great impact on cluster detection when the likelihood inference is intractable or unavailable.

1. Introduction

The spatial scan statistic is typically formulated under hypothesis testing problems with the null hypothesis that a disease rate is homogeneous in the entire region against the alternative hypothesis that the disease rate is elevated in a subregion. It has been successfully formulated under the framework of logistic linear models for Bernoulli or binomial data (Kulldorff and Nagarwalla, 1995) and loglinear models for Poisson data (Assuncao and Costa, 2006; Zhang and Lin, 2009). The spatial scan approach, which is carried out by a spatial scan statistic (Kulldorff, 1997; Tango and Takahashi, 2005), is popular and widely used in cluster detection and disease surveillance. It has been considered as an important and fundamental tool in spatial epidemiology. The spatial scan approach has been extended to detect clusters in multinomial data (Jung et al., 2010), normal data (Huang et al., 2009), and survival data (Bhatt and Tiwari, 2014; Huang et al., 2007). It has also been extended to account for spatial correlation (Loh and Zhu, 2007), overdispersion (Zhang et al., 2012), and inflated zeros (Cançado et al., 2014; de Lima et al., 2015, 2017). Previous spatial scan statistics are mostly formulated under the framework of likelihood ratio statistics. As the computation of a likelihood ratio statistic needs the exact distribution, the implementation of the previous spatial scan approach is difficult if the exact distribution is not provided or hard to compute.

The likelihood ratio-based spatial scan statistic has nice theoretical properties. By the Neyman–Pearson Lemma (Lehmann, 1986, P. 72), the uniformly most powerful (UMP) test can be formulated by the likelihood ratio statistic, indicating that likelihood ratio-based spatial scan statistics are powerful in detecting spatial clusters. It is not claimed by the Neyman–Pearson Lemma that the likelihood ratio statistic can dominate any other test statistic. We may have other tests which are as powerful as the likelihood ratio test. An example is the well-known t-test in linear models, which provides a uniformly

* Correspondence to: University of International Business and Economics, School of Statistics, #730, Chengxin Building, No. 10, Huixin Dongjie, Chaoyang District, Beijing 100029, China.

E-mail addresses: yliu@uibe.edu.cn (Y. Liu), liuyw@uibe.edu.cn (Y. Liu), tlzhang@purdue.edu (T. Zhang).

https://doi.org/10.1016/j.csda.2018.06.002
0167-9473/

most powerful unbiased (UMPU) test for the significance of regression coefficients (Lehmann, 1986, P. 397). As the t-statistic becomes the Wald statistic in linear regression, we study the Wald-based spatial scan approach in this article.

The formulation of Wald-based spatial scan statistics is consistent with output formats of general statistical models. If the distribution of a response variable is modeled by a linear function of explanatory variables, then any valid fitting procedure should provide estimates of linear coefficients and their variance–covariance matrix. A set of Wald statistics (Wald, 1943) is basically used to assess the significance of individual linear coefficients. A Wald-based spatial scan statistic can be formulated if we can transform spatial clusters into explanatory variables. Since the derivation can be based on any estimation methods, the Wald-based spatial scan statistic provides an important option if the computation of the likelihood ratio statistic is difficult or even impossible. Although initial ideas for Poisson data can be traced back (Zhang and Lin, 2009, 2013), the formal approach to statistical models beyond GLMs (generalized linear models) has not been investigated, which motivates the present research.

The proposed approach is important in extension and generalization of the spatial scan test for cluster detection. Note that Kulldorff's spatial scan statistic (Kulldorff, 1997) is constructed via a likelihood ratio statistic in a Bernoulli or a Poisson model. It cannot be used if the exact distribution of data is not provided or intractable. Tango and Takahashi's flexibly shaped spatial scan statistic (Tango and Takahashi, 2005) is also constructed via a likelihood ratio statistic. It faces the same problem if the likelihood function is not provided or intractable. An obvious and important example is the construction of the spatial scan statistic in the quasi-Poisson model (McCullagh, 1983). As the variability of the disease count exceeds the corresponding value provided by the Poisson model, disregarding the presence of overdispersion in the quasi-Poisson model may lead to an inflation of type I error probabilities (Zhang et al., 2012). This phenomenon is often termed as the overdispersion problem in GLMs. Since the exact distribution is usually not specified, the likelihood ratio statistic is generally not well-defined. To solve the problem, one can introduce a Gamma distribution for overdispersion in the quasi-Poisson model. This may induce a likelihood ratio-based quasi-Poisson spatial scan statistic, but the derivation of a likelihood ratio-based spatial scan statistic is hard if a normal prior is utilized. If a Wald-based approach is used, then we can address the difference between the choices of the Gamma and the normal distributions for overdispersion. In addition, the proposed Wald-based spatial scan approach bridges cluster detection and linear functions of statistical models. It can be easily combined with well-known statistical approaches when response and explanatory variables are involved.

The article is organized as follows. In Section 2, we briefly review the likelihood ratio-based spatial scan statistics. In Section 3, we propose the Wald-based spatial scan approach, which also contains its specifications to a few important statistical models, such as the negative binomial, the quasi-binomial, and the quasi-Poisson models. Note that they are not exponential family distributions. These examples indicate that the spatial scan test can still be used if the model is not a GLM. In Section 4, we numerically evaluate the properties of our Wald-based spatial scan statistic in comparison with the likelihood ratio-based spatial scan statistic. In Section 5, we provide a discussion.

2. Likelihood ratio-based spatial scan statistic

The scan approach was originally developed for one dimensional point process (Naus, 1965). By a likelihood ratio-based method, Kulldorff (1997) extended it to cluster detection for two-dimensional aggregated unit data when the response follows Poisson or Bernoulli distributions. Kulldorff's scan approach was later extended to other distributions. In order to understand the impact of our Wald-based spatial scan statistic, it is important to review the likelihood ratio-based spatial scan statistic at the beginning. Here we only review the approach to the Poisson data.

Suppose a study area has been partitioned into m spatial units. Each has an at-risk population size and a number of case counts. Let Y_i be the count, y_i be the observed count, and n_i be the at-risk population size in unit i, for $i = 1, \ldots, m$. Let $\mathcal{C}$ be the collection of cluster candidates. For a selected $C \in \mathcal{C}$, let $Y = \sum_{i=1}^{m} Y_i$, $y = \sum_{i=1}^{m} y_i$, $n = \sum_{i=1}^{m} n_i$, $Y_C = \sum_{i\in C} Y_i$, $y_C = \sum_{i\in C} y_i$, $n_C = \sum_{i\in C} n_i$, $Y_{\bar{C}} = \sum_{i\in \bar{C}} Y_i$, $y_{\bar{C}} = \sum_{i\in \bar{C}} y_i$, and $n_{\bar{C}} = \sum_{i\in \bar{C}} n_i$. Then, y, y_C and $y_{\bar{C}}$ are the observed values of Y, Y_C and $Y_{\bar{C}}$, respectively. Consider the test for the null hypothesis as

$$Y_i \sim Poisson(\theta_0 n_i),\ i = 1, \ldots, m, \tag{1}$$

against the alternative hypothesis for hot spot clusters as

$$Y_i \sim Poisson(\theta_0 n_i),\ i \in \bar{C};\ \text{or } Y_i \sim Poisson(\theta_C n_i),\ i \in C \in \mathcal{C},\ \theta_C > \theta_0. \tag{2}$$

The likelihood function is

$$L_C(\theta_0, \theta_C) = \left(\prod_{i=1}^{m} \frac{n_i^{Y_i}}{Y_i!}\right)\left(\prod_{i\in C} \theta_C^{Y_i} e^{-\theta_C n_i}\right)\left(\prod_{i\in \bar{C}} \theta_0^{Y_i} e^{-\theta_0 n_i}\right). \tag{3}$$

The likelihood ratio statistic is

$$\Lambda_C = \frac{\max_{\theta_C>\theta_0} L_C(\theta_0, \theta_C)}{\max_{\theta_C=\theta_0} L_C(\theta_0, \theta_C)} = \left(\frac{Y_C/n_C}{Y/n}\right)^{Y_C}\left(\frac{Y_{\bar{C}}/n_{\bar{C}}}{Y/n}\right)^{Y_{\bar{C}}}, \tag{4}$$

Please cite this article in press as: Liu Y., et al., Wald-based spatial scan statistics for cluster detection. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.06.002.