LOG-LINEAR, LOGISTIC MODEL FITTING AND LOCAL SCORE STATISTICS FOR CLUSTER DETECTION WITH COVARIATE ADJUSTMENTS

Hock Peng Chan*

Department of Statistics and Applied Probability, National University of Singapore, Singapore

I-Ping Tu

Institute of Statistical Science, Academia Sinica, Taiwan

SUMMARY

The standard method for p-value computation of spatial scan statistics, with adjustments for covariate effects, is to conduct Monte Carlo simulations with these effects estimated under the null hypothesis of no clustering. However when the covariates are geographically unbalanced, the proposed Monte Carlo p-value estimates are too conservative, with corresponding loss of power, due to excessive adjustments for confounding between covariates and location. We show that the use of an alternative procedure that involves local score statistics, with parameters fitted on a log-linear or logistic model, addresses this problem. We also discuss extensions of the procedure when there are multiple or continuous covariates.

KEY WORDS: cluster detection; local score statistics; log-linear model; logistic model; scan statistic

1. INTRODUCTION

The spatial scan statistic is a popular tool for analyzing epidemiological datasets for evidence of local or "hot-spot" clusters with unusually high or low disease rates, and is widely implemented through the SaTScan software. When covariates are available, the SaTScan program requires the user to estimate the coefficients of these covariates using an external regression software, then conduct Monte Carlo simulations with these estimates plugged into the model for p-value computations, see [1] pp.17–20. The Monte Carlo method in itself is not a problem but the use of a likelihood ratio based on the Poisson model penalizes counfounding between covariate and location effects too heavily. To obtain the appropriate likelihood ratio test statistic, we extend the Poisson model into a log-linear model that incorporates estimation of the covariate effects. For run-time considerations when evaluating p-values via Monte Carlo methods, we also propose the use of approximating local score statistics. When the disease rates and covariates are both affected by location, the scenario that leads us to include these covariates in the first place, these modifications will result in an increase in the power of the spatial scan statistic. We describe the procedure in Section 2, and use qq-plots to illustrate why we are able to address the drawbacks of the spatial scan statistic the way it is currently being used. In Section 3, we apply our methodology on a brain cancer dataset and discuss extensions to multiple or continuous covariates and the logistic model.

2. MONTE CARLO RISK-ADJUSTED P-VALUES

2.1 *Test statistics based on the Poisson model*

The Monte Carlo method for computing p-values of the spatial scan statistic that follows from [3] has been popularized by the SaTScan software. Let $t_1, \ldots, t_I$ be location co-ordinate vectors, representing perhaps the centroids of I counties or census tracts. Let n_i be the number of subjects assigned to t_i, with m_i of them having a specified medical condition, for example a disease, that is suspected to have a local environmental source. Then $n = \sum_{i=1}^{I} n_i$ is the total number of subjects, $m = \sum_{i=1}^{I} m_i$ the total number of disease cases, and $\widehat{p} = m/n$ the overall proportion of disease cases. For any $A \subset \{1, \ldots, I\}$, we let $n_A = \sum_{i \in A} n_i$, $m_A = \sum_{i \in A} m_i$ and $\widehat{p}_A = m_A/n_A$. We shall also use the symbol B to denote the complement of A, that is $B = \{i : 1 \leq i \leq I, i \notin A\}$. Typically, we are interested in sets A such that $\{t_i : i \in A\}$ are clustered.

Let p_A be the disease rate for subjects in A (more precisely, for subjects assigned to t_i, $i \in A$). The generalized likelihood ratio (GLR) statistic for testing H_0: $p_A = p_B$ against H_1: $p_A \neq p_B$ is

$$\Big(\frac{\widehat{p}_A}{\widehat{p}}\Big)^{m_A}\Big(\frac{1-\widehat{p}_A}{1-\widehat{p}}\Big)^{n_A-m_A}\Big(\frac{\widehat{p}_B}{\widehat{p}}\Big)^{m_B}\Big(\frac{1-\widehat{p}_B}{1-\widehat{p}}\Big)^{n_B-m_B}.$$

When we have large population sizes and the disease rate is low, Poisson approximation of the binomial distribution can be applied, and the corresponding GLR test statistic is

$$L_A^* = \Big(\frac{m_A}{n_A\widehat{p}}\Big)^{m_A}\Big(\frac{m_B}{n_B\widehat{p}}\Big)^{m_B}. \tag{2.1}$$

Under H_0,

$$2\log L_A^* \xrightarrow{d} \chi_1^2 \text{ as } n_A, n_B \to \infty,$$

where $\xrightarrow{d}$ denotes convergence in distribution, and χ_1^2 denotes the chi-square distribution with one degree of freedom (d.f.).

When we have a class of scanning sets $\mathcal{A}$, the spatial scan statistic $M = \max_{A \in \mathcal{A}} 2\log L_A^*$ is often used to determine significance. Unlike $2\log L_A^*$, the asymptotic distribution of M cannot be characterized in a simple manner, and randomized permutation tests, see [2], are often used to evaluate p-values of M. Let K be the designated number of simulation runs. For each $1 \leq k \leq K$, the m disease cases are permuted randomly among the n subjects and the spatial scan statistic $M^{(k)}$, corresponding to this random permutation, is computed. The Monte Carlo p-value of M is then

$$\frac{1 + (\#\{k : M^{(k)} \geq M\})}{1+K}, \tag{2.2}$$

where $\#$ denotes the cardinality or size of a finite set.

Adjusting for the effects of covariates is important, see [3, 4, 5]. When covariates are present, permutation of the disease cases together with their covariates is inappropriate as it destroys the covariate-location structure of the dataset. The SaTScan software uses the following Poisson model approach. Let there be a single covariate with J values or categories. Let n_{ij} be the number of

subjects at location i that are in category j, $1 \le j \le J$, with m_{ij} of them having the disease. Let $m_{\cdot j} = \sum_{i=1}^{I} m_{ij}$ be the pooled number of disease cases and $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$ the number of subjects in category j. The estimated disease rate of category j subjects is $\widehat{p}_{\cdot j} = m_{\cdot j}/n_{\cdot j}$. Then the risk estimate at location i is $\widehat{\lambda}_i = \sum_{j=1}^{J} n_{ij}\widehat{p}_{\cdot j}$. Let $\widehat{\lambda}_A = \sum_{i \in A} \widehat{\lambda}_i$ be the total estimated risks for the subjects in A. To obtain the GLR statistic $\widetilde{L}_A$ for testing the null hypothesis of no clustering in A, the suggested method is to replace $n_A\widehat{p}$ by $\widehat{\lambda}_A$ and $n_B\widehat{p}$ by $\widehat{\lambda}_B$ in (2.1). Hence

$$\widetilde{L}_A = \Big(\frac{m_A}{\widehat{\lambda}_A}\Big)^{m_A} \Big(\frac{m_B}{\widehat{\lambda}_B}\Big)^{m_B}. \tag{2.3}$$

For a class of scanning sets $\mathcal{A}$, the scan statistic is $\widetilde{M} = \max_{A \in \mathcal{A}} 2\log \widetilde{L}_A$. To estimate the p-value of $\widetilde{M}$, generate in the kth simulation run, for $1 \le k \le K$,

$$(m_1^{(k)}, \ldots, m_I^{(k)}) \sim \text{Multinomial}(m; \widehat{\lambda}_1/m, \ldots, \widehat{\lambda}_I/m),$$

and use it to compute

$$\widetilde{L}_A^{(k)} = \Big(\frac{m_A^{(k)}}{\widehat{\lambda}_A}\Big)^{m_A^{(k)}} \Big(\frac{m_B^{(k)}}{\widehat{\lambda}_B}\Big)^{m_B^{(k)}}. \tag{2.4}$$

The p-value of $\widetilde{M}$ is then as in (2.2), with $\widetilde{M}$ replacing M and $\widetilde{M}^{(k)} = \max_{A \in \mathcal{A}} 2\log \widetilde{L}_A^{(k)}$ replacing $M^{(k)}$.

2.2 *Test statistics based on the log-linear model*

We do not have exact correspondence between the expressions of $\widetilde{L}_A$ and $\widetilde{L}_A^{(k)}$ and this causes problems when evaluating p-values. Consider for example a dummy covariate that divides up the population into categories 1 and 2, and let there be equal number of subjects both inside and outside a scanning set A. If there are more category 1 subjects in A compared to outside A and coincidentally the observed disease rate is higher in A, then we tend to assign higher risks for subjects in A. When running the simulations, we do so with risk parameters $\widehat{\lambda}_A > \widehat{\lambda}_B$ even though the underlying risks are the same for both inside and outside A, and the significance of observing the m_A cases in A is unfairly reduced.

For this reason, we advocate to incorporate the risk estimation procedure as parameter estimation within a larger log-linear model. Let α_i be the parameter associated with the ith location and β_j the parameter associated with the jth category of subjects. Consider the additive log-linear model

$$\begin{aligned} &m_{ij} \sim \text{Poisson}(\lambda_{ij}), \text{ where } \log(\lambda_{ij}/n_{ij}) = \alpha_i + \beta_1 x_{j1} + \cdots \beta_J x_{jJ}, \\ &\text{where } x_{jk} = 1 \text{ if } j = k \text{ and } x_{jk} = 0 \text{ otherwise.} \end{aligned} \tag{2.5}$$

Under the null model of no location effects, $\alpha_1 = \cdots = \alpha_I = 0$, and the model reduces to

$$m_{\cdot j} \sim \text{Poisson}(\lambda_{\cdot j}), \text{ where } \log(\lambda_{\cdot j}/n_{\cdot j}) = \beta_1 x_{j1} + \cdots + \beta_J x_{jJ}, \tag{2.6}$$

with maximum likelihood estimates (MLE) $\widehat{\beta}_j = \log \widehat{p}_{\cdot j} = \log(m_{\cdot j}/n_{\cdot j})$. Under the alternative model that disease rates may differ for subjects inside A versus outside A, we impose the constraints $\alpha_i = \alpha_A$

for $i \in A$ and $\alpha_i = \alpha_B$ for $i \notin A$, with α_A, α_B unspecified. Hence for all $1 \le j \le J$,

$$\begin{aligned} m_{Aj} &\sim \text{Poisson}(\lambda_{Aj}), \text{ with } \log(\lambda_{Aj}/n_{Aj}) = \alpha_A + \beta_1 x_{j1} + \cdots + \beta_J x_{jJ}, \\ m_{Bj} &\sim \text{Poisson}(\lambda_{Bj}), \text{ with } \log(\lambda_{Bj}/n_{Bj}) = \alpha_B + \beta_1 x_{j1} + \cdots + \beta_J x_{jJ}, \end{aligned} \tag{2.7}$$

where $m_{Aj} = \sum_{i \in A} m_{ij}$, $n_{Aj} = \sum_{i \in A} n_{ij}$ and similarly for m_{Bj}, n_{Bj}. Let L_A be the generalized likelihood ratio between models (2.7) and (2.6). As the difference in the number of constraints between the null and alternative models is one,

$$2 \log L_A \xrightarrow{d} \chi^2_1.$$

This follows from standard statistical theory discussed in for example [6, 7].

The log-linear model, in the context of spatial cluster detection, was discussed by Jung [8] as a special case of the generalized linear models (see also [9]), essentially with a single response at each location. We extended [8] by fitting multiple responses at each location, one for each category. Separating the responses allows us to avoid over-dispersion due to mixing together of different population types, see [10, 11] for related discussions.

2.3 *Local score statistics*

It is computationally very expensive to fit a log-linear model for each $A \in \mathcal{A}$ in order to compute $M = \max_{A \in \mathcal{A}} 2 \log L_A$ when $\#\mathcal{A}$ is large. For large $\#\mathcal{A}$ situations, we propose the use of local score statistics as a computationally attractive alternative to the direct computation of $2 \log L_A$. Let $\mathbf{I}_J$ be the $J \times J$ identity matrix and let

$$\mathbf{X} = \begin{pmatrix} \mathbf{I}_J \\ \vdots \\ \mathbf{I}_J \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} n_{11}\widehat{p}_{\cdot 1} & & & \\ & n_{12}\widehat{p}_{\cdot 2} & & \\ & & \ddots & \\ & & & n_{IJ}\widehat{p}_{\cdot J} \end{pmatrix}. \tag{2.8}$$

The matrix $\mathbf{X}$ is of size $IJ \times J$ while $\mathbf{W}$ is a diagonal matrix of size $IJ \times IJ$. Let

$$\mathbf{U} = \mathbf{I}_{IJ} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}' \text{ and } \mathbf{V} = \mathbf{U}\mathbf{W} = \mathbf{W} - \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}. \tag{2.9}$$

Note that the matrices to be inverted in (2.9) are of size $J \times J$ and not the much larger $IJ \times IJ$. Let $\mathbf{m} = (m_{11}, m_{12}, \ldots, m_{IJ})'$ and $\widehat{\mathbf{m}} = (n_{11}\widehat{p}_{\cdot 1}, n_{12}\widehat{p}_{\cdot 2}, \ldots, n_{IJ}\widehat{p}_{\cdot J})'$. For Monte Carlo p-value computation, we simulate, for each $1 \le k \le K$,

$$\mathbf{m}^{(k)} = (m_{11}^{(k)}, m_{12}^{(k)}, \ldots, m_{IJ}^{(k)})' \sim \text{Multinomial}(m, \widehat{\mathbf{m}}/m). \tag{2.10}$$

Let $L_A^{(k)}$ be the GLR between (2.7) and (2.6), for observations $\mathbf{m}^{(k)}$.

Theorem 1. Consider the additive log-linear model (2.5). Let $v_A = (v_{A11}, v_{A12}, \ldots, v_{AIJ})'$, with

$$v_{Aij} = \begin{cases} 1 & \text{if } i \in A, \\ 0 & \text{otherwise.} \end{cases}$$

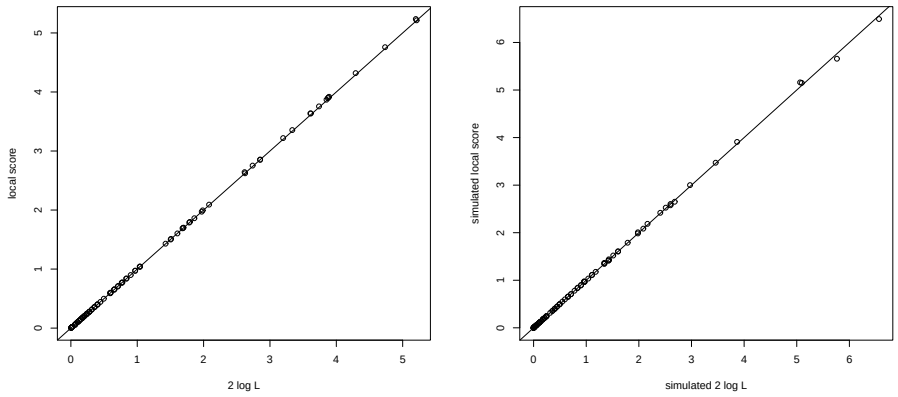

Figure 1: **Log-linear model**. Scatter-plots of s_A against $2\log L_A$ (left) and $s_A^{(k)}$ against $2\log L_A^{(k)}$ (right).

(a) Under the null model (2.6), as $n_{Aj} \to \infty$ and $n_{Bj} \to \infty$ for some j,

$$2\log L_A - s_A \xrightarrow{d} 0, \text{ where } s_A = \frac{|v_A'(\mathbf{m} - \widehat{\mathbf{m}})|^2}{v_A'\mathbf{V}v_A}. \tag{2.11}$$

(b) Let $\mathbf{m}^{(k)}$ be generated as in (2.10). Then as $n_{Aj} \to \infty$ and $n_{Bj} \to \infty$ for some j,

$$2\log L_A^{(k)} - s_A^{(k)} \xrightarrow{d} 0, \text{ where } s_A^{(k)} = \frac{|v_A'\mathbf{U}(\mathbf{m}^{(k)} - \widehat{\mathbf{m}})|^2}{v_A'\mathbf{V}v_A}. \tag{2.12}$$

In view of (2.11), we replace $2\log L_A$ by the local score statistic s_A and compute $M = \max_{A\in\mathcal{A}} s_A$. For purpose of p-value evaluation, we compare M against $M^{(k)} = \max_{A\in\mathcal{A}} s_A^{(k)}$. Notice that there is no need to fit a log-linear model under (2.6), even though the theory is based on the log-linear model, as the parameter estimates $\widehat{p}_{\cdot j} = m_{\cdot j}/n_{\cdot j}$ are computed directly. Iterative least-squares for estimation of fitted-values is however needed in Section 3.2.

In Figure 1, we check that both s_A and $s_A^{(k)}$ approximate their respective likelihood ratios $2\log L_A$ and $2\log L_A^{(k)}$ very well for $m = 100$ cases. The plots are constructed for the scenario $I = 10$, $J = 3$, with $A = \{1, 2, 3, 4, 5\}$, $(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}) = (6, 3, 1)$ for $1 \le i \le 5$ and $(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}) = (3, 6, 1)$ for $6 \le i \le 10$.

3. NUMERICAL STUDIES AND EXTENSIONS

3.1 *New Mexico brain cancer dataset*

This is a dataset downloaded from the SaTScan web-site and was studied in [12]. The total population sizes and population representation by race were taken from a 1973 census. In the dataset,

Location	White	Black	Other	Population
1. Los Alamos	98.5%	0.4%	1.0%	15,315
2. McKinley	36.1%	0.8%	63.0%	46,826
3. Rio Arriba	88.6%	0.2%	11.2%	27,339
4. San Juan	65.3%	0.5%	34.2%	58,718
5. Valencia	84.7%	0.5%	14.8%	43,192
6. Sandoval	66.2%	0.4%	33.5%	23,858

Table 1: Population break-down and population sizes of six counties in northwestern New Mexico.

RR	1	1.1	1.2	1.3	1.4	1.5
Poisson	0.048	0.062	0.210	0.497	0.849	0.971
Local	0.048	0.086	0.387	0.706	0.961	0.999

Table 2: P-values and detection power comparisons after adjusting for race only.

the races were coded as 1="white", 2="black" and 3="other". A total of $m = 1175$ brain cancer cases were recorded and the proportion of brain cancer cases within each race were $(p_{\cdot 1}, p_{\cdot 2}, p_{\cdot 3}) = (1.12, 0.42, 0.56) \times 10^{-3}$. The population break-down by race of the (then) $I = 32$ counties in New Mexico were also provided, see Table 1 for the population break-down of six counties in northwestern New Mexico. Let n_{ij} be the number of subjects of race j at location i, with m_{ij} of them having brain cancer. We want to test if there exists a cluster of counties with brain cancer rates that are significantly higher than the other counties, after adjusting for race. For the numerical exercises in this section, we selected $K = 99$ for the computation of Monte Carlo p-values.

Let d_{ih} be the distance between t_i and t_h, where t_i is the location co-ordinate vector of the ith county. Let $0 = d_{i(1)} \leq \cdots \leq d_{i(I)}$ be the ordered values of $d_{i1}, \ldots, d_{iI}$. Let $A_{ih} = \{j : d_{ij} \leq d_{i(h)}\}$ be the cluster of h counties closest to county i and let h_i be the largest h such that the population size in A_{ih} does not exceed half the total population. We test for the presence of spatial effects by considering the collection of clusters

$$\mathcal{A} = \{A_{ih} : 1 \leq i \leq I, 1 \leq h \leq h_i\}. \tag{3.1}$$

By using local score statistics of the log-linear model, we obtained a maximum score $M = 13.66$ and a marginally significant Monte Carlo p-value of 0.03. When we applied the Poisson model described in Section 2.1, we obtained scan statistic $\widetilde{M} = 13.57$ and the Monte Carlo p-value was again 0.03.

For detection power comparisons, we chose the set A_{46}, the cluster of six counties listed in Table 1, which have in general higher proportions of "other" compared to the other twenty-six counties. We simulated the disease cases using disease rates $(p_{i1}, p_{i2}, p_{i3}) = (1.12, 0.42, 0.56) \times 10^{-3}$ for $i \notin A_{46}$ and $(p_{i1}, p_{i2}, p_{i3}) = \text{RR} \times (1.12, 0.42, 0.56) \times 10^{-3}$ for $i \in A_{46}$, RR denoting relative risk.

One thousand experiments were conduced for each value of $\text{RR} = 1, 1.1, 1.2, 1.3, 1.4, 1.5$. For each

RR	1	1.1	1.2	1.3	1.4	1.5
Poisson	0.037	0.049	0.096	0.281	0.557	0.853
Local	0.042	0.060	0.138	0.357	0.673	0.919

Table 3: P-values and detection power comparisons after adjusting for both race and age.

experiment, we computed the maximum local score M based on the log-linear model, and the Monte Carlo maximum local scores $M^{(k)}$. We also computed the scan statistic $\widetilde{M}$ and Monte Carlo scan statistics $\widetilde{M}^{(k)}$, based on the Poisson model. We rejected the null hypothesis of no cluster effects if the Monte Carlo p-value does not exceed 0.05. We see from Table 2 for the column RR = 1 that both methods give reasonably accurate p-values. However the "local" approach is much better compared to the "Poisson" approach at detecting an actual difference in disease rates, and can achieve an increase in detection power of more than 20%.

3.2 *Multiple and continuous covariates*

When there are multiple covariates, each taking finitely many values or grouped into finitely many categories, we can in principle treat them as a single covariate with a large number of categories. An alternative is to divide-up the subjects in location i into r_i broad categories, and model

$$m_{ir} \sim \text{Poisson}(\lambda_{ir}), \text{ where } \log(\lambda_{ir}/n_{ir}) = \alpha_i + \sum_{j=1}^{J} \beta_j x_{irj}, \text{ for } 1 \le r \le r_i, \tag{3.2}$$

where n_{ir} is the number of subjects and m_{ir} the number of cases in category r. The use of local score statistics is justfied by the following extension of Theorem 1.

Theorem 2. Let $A \in \mathcal{A}$ and consider the model (3.2) with $\alpha_i = \alpha_A$ for all $i \in A$, $\alpha_i = \alpha_B$ for all $i \in B$, for some α_A, α_B. Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix}, \text{ where } \mathbf{Y}_i = \begin{pmatrix} x_{i11} & \cdots & x_{i1J} \\ \vdots & \ddots & \vdots \\ x_{ir_i1} & \cdots & x_{ir_iJ} \end{pmatrix}, \text{ and let } \mathbf{W} = \begin{pmatrix} \widehat{\lambda}_{11} & & & \\ & \widehat{\lambda}_{12} & & \\ & & \ddots & \\ & & & \widehat{\lambda}_{Ir_I} \end{pmatrix},$$

where $\widehat{\lambda}_{ij}$ are the fitted-values under the null model $\alpha_A = \alpha_B$. Let $\mathbf{U}$ and $\mathbf{V}$ be given in (2.9). Assume that the population is distributed such that $\alpha_A - \alpha_B$ is consistently estimated by its MLE as $n \to \infty$.

(a) Under the null model $\alpha_A = \alpha_B$, (2.11) holds.

(b) Let $m^{(k)}$ be given in (2.10), with $\widehat{\mathbf{m}} = (\widehat{\lambda}_{11}, \widehat{\lambda}_{12}, \ldots, \widehat{\lambda}_{Ir_I})$. Then (2.12) holds.

For example, if we would like to adjust for both age and race in the New Mexico brain cancer dataset, we can model ninety-six responses, three responses at each location, corresponding to "white", "black" and "other", using $J = 4$ covariates. The first three covariates are used for modelling possible

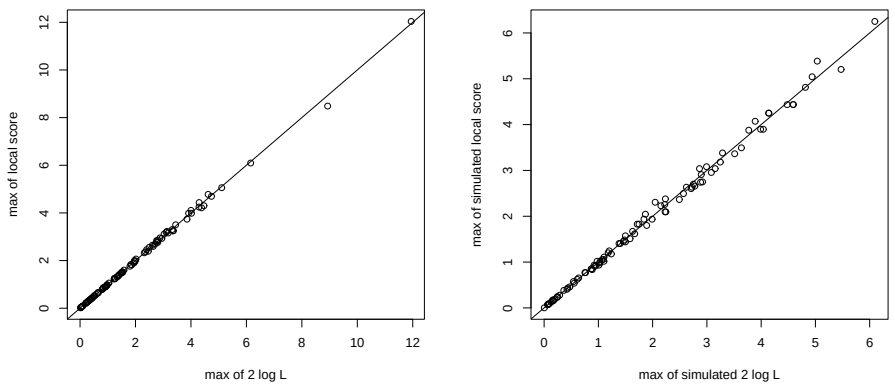

Figure 2: **Logistic model**. Scatter-plots of $\max_{A\in\mathcal{A}} s_A$ against $\max_{A\in\mathcal{A}} 2\log L_A$ (left) and $\max_{A\in\mathcal{A}} s_A^{(k)}$ against $\max_{A\in\mathcal{A}} 2\log L_A^{(k)}$ (right).

differences in disease rates among the three racial groups while the fourth covariate x_{ir4} is the average age of race r subjects at location i. The covariate and weight matrices are

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & x_{114} \\ 0 & 1 & 0 & x_{124} \\ 0 & 0 & 1 & x_{134} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & x_{I14} \\ 0 & 1 & 0 & x_{I24} \\ 0 & 0 & 1 & x_{I34} \end{pmatrix} \text{ and } \mathbf{W} = \begin{pmatrix} \widehat{\lambda}_{11} & & & & \\ & \widehat{\lambda}_{12} & & & \\ & & \ddots & & \\ & & & \widehat{\lambda}_{I3} & \end{pmatrix} \text{ with } I = 32.$$

The entries of the diagonal matrix $\mathbf{W}$ are fitted-values of (3.2) under the null model $\alpha_1 = \cdots = \alpha_I (= 0)$.

Let $\mathcal{A}$ be the collection of clusters given in (3.1). The maximum local score of the log-linear model was $M = 15.23$ and its Monte Carlo p-value 0.02. For the Poisson model, we estimate the risk in the ith county by $\widehat{\lambda}_i = \widehat{\lambda}_{i1} + \widehat{\lambda}_{i2} + \widehat{\lambda}_{i3}$. Then the risk estimates $\widehat{\lambda}_A = \sum_{i\in A} \widehat{\lambda}_i$ and $\widehat{\lambda}_B = \sum_{i\in B} \widehat{\lambda}_i$. Applying the formulas (2.3) and (2.4), we obtained scan statistic $\widetilde{M} = 14.80$ and Monte Carlo p-value 0.02.

As in Section 3.1, we conducted a power comparison study, but this time adjusting for both age and race instead of for race alone. The detection power was found to be larger for the local score statistic approach, see Table 3. The brain cancer cases were generated using (3.2), with parameters $\beta_1 = -7.36$, $\beta_2 = -8.26$, $\beta_3 = -7.96$, $\beta_4 = 0.0872$, $\alpha_i = 0$ for $i \notin A_{46}$ and $\alpha_i = \log \text{RR}$ for $i \in A_{46}$.

3.3 *Logistic model*

For smaller population sizes or when the disease rates are relatively large, we can consider applying

a logistic model. Let us consider the simpler case of a single covariate dividing the population into J categories. Let m_{ij} and n_{ij} be the number of disease cases and population size respectively in category j at center i. We model

$$m_{ij} \sim \text{Binomial}(n_{ij}, p_{ij}), \text{ where } \log\Big(\frac{p_{ij}}{1-p_{ij}}\Big) = \alpha_i + \beta_j, \qquad 1 \le i \le I, 1 \le j \le J. \tag{3.3}$$

Let $m_{\cdot j} = \sum_{i=1}^{I} m_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$. Under the null hypothesis of no location effects, $\alpha_1 = \cdots = \alpha_I = 0$, and we have the reduced model

$$m_{\cdot j} \sim \text{Binomial}(n_{\cdot j}, p_{\cdot j}), \text{ where } p_{\cdot j} = e^{\beta_j}/(1+e^{\beta_j}). \tag{3.4}$$

Let $m_{Aj} = \sum_{i\in A} m_{ij}$, $n_{Aj} = \sum_{i\in A} n_{ij}$ and similarly when A is replaced by B. If the disease rate differs for subjects inside A versus outside A, then analogous to (2.7),

$$\begin{aligned} m_{Aj} &\sim \text{Binomial}(n_{Aj}, p_{Aj}), \text{ with } p_{Aj} = e^{\alpha_A+\beta_j}/(1+e^{\alpha_A+\beta_j}), \\ m_{Bj} &\sim \text{Binomial}(n_{Bj}, p_{Bj}), \text{ with } p_{Bj} = e^{\alpha_B+\beta_j}/(1+e^{\alpha_B+\beta_j}). \end{aligned} \tag{3.5}$$

For p-value estimation, we generate in the kth simulation run,

$$m_{ij}^{(k)} \sim \text{Binomial}(n_{ij}, \widehat{p}_{\cdot j}), \text{ where } \widehat{p}_{\cdot j} = m_{\cdot j}/n_{\cdot j}, \text{ for } 1 \le i \le I, 1 \le j \le J. \tag{3.6}$$

The local score statistics s_A and $s_A^{(k)}$ described in (2.8), (2.9), (2.11) and (2.12) can again be used, but for the logistic model, which uses the binomial instead of the Poisson model, the diagonal weight matrix is

$$\mathbf{W} = \begin{pmatrix} n_{11}\widehat{p}_{\cdot 1}(1-\widehat{p}_{\cdot 1}) & & & \\ & n_{12}\widehat{p}_{\cdot 2}(1-\widehat{p}_{\cdot 2}) & & \\ & & \ddots & \\ & & & n_{IJ}\widehat{p}_{\cdot J}(1-\widehat{p}_{\cdot J}) \end{pmatrix}. \tag{3.7}$$

Let L_A and $L_A^{(k)}$ be the GLR between (3.5) and (3.4) for observations $\mathbf{m}$ and $\mathbf{m}^{(k)}$ respectively.

Theorem 3. Let $\mathbf{U}$, $\mathbf{V}$ be defined in (2.9) with $\mathbf{W}$ given in (3.7).

(a) The relation (2.11) holds under (3.4) when $n_{Aj} \to \infty$ and $n_{Bj} \to \infty$ for some j.

(b) The relation (2.12) holds under (3.3), with $\mathbf{m}^{(k)}$ generated in (3.6), when $n_{Aj} \to \infty$ and $n_{Bj} \to \infty$ for some j.

We see from Figure 2 that the local score statistics and simulated local score statistics are good approximations of the test statistics L_A and $L_A^{(k)}$, as indicated in Theorem 3. Figure 2 is generated using the reduced model (3.4) with $I = 3$, $J = 2$, $n_{i1} = 100i$, $n_{i2} = 300 + 100i$ for $i = 1, 2, 3$, $\beta_1 = -1$, $\beta_2 = -2$ and $\mathcal{A} = \{\{1\}, \{2\}, \{3\}\}$.

4. DISCUSSION

In this paper, we show how we can capture the location-covariate dependencies in an epidemiological dataset more accurately by extending the commonly used Poisson model to a larger log-linear model. We also ease the computational burden of dealing with the log-linear model by proposing local score approximations. The log-linear model improves upon the Poisson model by not over-correcting for unbalanced covariates and we do not expect any power difference between the two models when the spatial effect is on a cluster with balanced covariates (between inside and outside the cluster). However, as backed up by our numerical studies in Sections 3.1 and 3.2, the power gain when applying the log-linear model can be substantial when the spatial effect occurs on a cluster with unbalanced covariates.

In the New Mexico dataset, the population size is large at each location. In some datasets, the geographical resolution is very fine so that I is large and n_i, $1 \leq i \leq I$ relatively small. The local score approximations still hold as long as we restrict ourselves to scanning sets A such that the number of subjects in both A and its complement B is large. We are unlikely to be able to detect spatial clustering of a very small population base so imposing this restriction will not seriously affect the detection capabilities of the spatial scan statistic. Moreover, imposing this restriction makes the computation of the scan statistic more manageable as the size of $\mathcal{A}$ is reduced. Notice that in Theorem 1, we only require that the population size of one category be large in both inside and outside A. Hence we can apply local score approximations even when we have a category of subjects with a very small population size. For example in Figure 1, the approximations were fine even though the risk for the third category of subjects at each location was very small.

Our purpose in Section 3.3, where we briefly discuss the logistic model, is to show that when dealing with a different model, what changes in the expression of the local score statistic is the form of the weight matrix $\mathbf{W}$. Hence we have chosen simple categorical covariates when discussing this model. Extensions of local score approximations to general covariate matrices $\mathbf{X}$ encoding multiple and continuous covariates, as discussed in Section 3.2, apply to logistic models as well.

APPENDIX: PROOF OF THEOREMS 1–3

We only prove Theorems 2 and 3 since Theorem 1 is a special case of Theorem 2. The proofs of Theorems 2 and 3 are based on likelihood equations of the log-linear and logistic models respectively. See Section 8.6.2 of [13] for the derivation of these equations.

A.1 *Proof of Theorem* 2: *Log-linear model*

We relabel the subscripts and consider a general covariate matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1J} \\ \vdots & & \vdots \\ x_{u1} & \cdots & x_{uJ} \end{pmatrix}, \text{ where } u = \sum_{i=1}^{I} r_i,$$

see (3.2). Let $v_i = 1$ for $i \leq a$ and $v_i = 0$ for $i > a$ for some $1 \leq a < u$. The log-likelihood under the log-linear

model

$$m_i \sim \text{ Poisson}(\lambda_i), \text{ where } \log(\lambda_i/n_i) = \alpha v_i + \sum_{j=1}^{J} \beta_j x_{ij}, \text{ for } 1 \le i \le u,$$

is

$$\ell(\alpha, \beta_1, \ldots, \beta_J) = \sum_{i=1}^{u} \Big\{ m_i \Big(\alpha v_i + \sum_{j=1}^{J} \beta_j x_{ij} \Big) - n_i \exp\Big(\alpha v_i + \sum_{j=1}^{J} \beta_j x_{ij} \Big) \Big\} + \text{ constant.} \tag{A.1}$$

Let $\ell_j = \frac{d\ell}{d\beta_j}\Big|_{(0,\hat\beta)}$ (with $\beta_0 = \alpha$), and $\ell_{jk} = \frac{d^2\ell}{d\beta_j d\beta_k}\Big|_{(0,\hat\beta)}$. Since $(0, \widehat\beta)$ is the maximum likelihood when constrained on $\alpha = 0$,

$$\ell_1 = \cdots = \ell_J = 0. \tag{A.2}$$

Let $\widehat m_i = \widehat\lambda_i$. By (A.1),

$$\ell_0 = \sum_{i=1}^{a} (m_i - \widehat m_i), \quad \ell_{00} = -\sum_{i=1}^{a} \widehat m_i, \quad \ell_{0j} = -\sum_{i=1}^{a} x_{ij} \widehat m_i, \quad \ell_{jh} = -\sum_{i=1}^{u} x_{ij} x_{ih} \widehat m_i. \tag{A.3}$$

Let $\mathbf{v} = (v_1, \ldots, v_u)'$ and $\mathbf{W}$ a $u \times u$ diagonal matrix with entries $\widehat m_i$. Then by (A.2) and (A.3), the second-order expansion of $2[\ell(\alpha, \widehat\beta + \delta) - \ell(0, \widehat\beta)]$ is

$$Q(\alpha, \delta) = 2\ell_0 \alpha - (\alpha \mathbf{v} + \mathbf{X}\delta)' \mathbf{W} (\alpha \mathbf{v} + \mathbf{X}\delta). \tag{A.4}$$

For fixed α, $Q(\alpha, \delta)$ is maximized when $\delta = -\alpha (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{v}$, and hence by (2.9),

$$\sup_{\delta} Q(\alpha, \delta) = 2\ell_0 \alpha - \alpha^2 \mathbf{v}'\mathbf{V}\mathbf{v} \Rightarrow \sup_{\alpha,\delta} Q(\alpha, \delta) = \frac{\ell_0^2}{\mathbf{v}'\mathbf{V}\mathbf{v}}. \tag{A.5}$$

Theorem 1(a) follows from (A.5), since $2 \log L_A = \sup_{\alpha,\delta} Q(\alpha, \delta) + o_d(1)$.

Let $\ell^{(k)}$ be the log-likelihood with respect to simulated observations $\mathbf{m}^{(k)}$. Since $(0, \widehat\beta)$ is now the underlying parameters instead of the MLE, the partial derivatives of $\ell^{(k)}$ at $(0, \widehat\beta)$ satisfy

$$\ell_0^{(k)} = \sum_{i=1}^{a} (m_i^{(k)} - \widehat m_i), \quad \ell_j^{(k)} = \sum_{i=1}^{u} x_{ij} (m_i^{(k)} - \widehat m_i), \quad \ell_{ij}^{(k)} = \ell_{ij}. \tag{A.6}$$

It follows from (A.6) that

$$\begin{aligned} 2 \log L_A^{(k)} &= \sup_{\alpha,\delta} Q^{(k)}(\alpha, \delta) - \sup_{\delta} Q^{(k)}(0, \delta) + o_d(1), \qquad (A.7) \\ \text{where } Q^{(k)}(\alpha, \delta) &= 2(\alpha \mathbf{v} + \mathbf{X}\delta)'(\mathbf{m}^{(k)} - \widehat{\mathbf{m}}) - (\alpha \mathbf{v} + \mathbf{X}\delta)' \mathbf{W} (\alpha \mathbf{v} + \mathbf{X}\delta) \\ &= (\mathbf{m}^{(k)} - \widehat{\mathbf{m}})' \mathbf{W}^{-1} (\mathbf{m}^{(k)} - \widehat{\mathbf{m}}) \\ &\quad - [\alpha \mathbf{v} + \mathbf{X}\delta - \mathbf{W}^{-1}(\mathbf{m}^{(k)} - \widehat{\mathbf{m}})]' \mathbf{W} [\alpha \mathbf{v} + \mathbf{X}\delta - \mathbf{W}^{-1}(\mathbf{m}^{(k)} - \widehat{\mathbf{m}})]. \end{aligned}$$

Since $\alpha \mathbf{v} + \mathbf{X}\delta = (\mathbf{v}\ \mathbf{X}) \binom{\alpha}{\delta}$, $Q^{(k)}(\alpha, \delta)$ is maximized when

$$\binom{\alpha}{\delta} = [(\mathbf{v}\ \mathbf{X})' \mathbf{W} (\mathbf{v}\ \mathbf{X})]^{-1} (\mathbf{v}\ \mathbf{X})' (\mathbf{m}^{(k)} - \widehat{\mathbf{m}}),$$

and

$$\sup_{\alpha,\delta} Q^{(k)}(\alpha, \delta) = (\mathbf{m}^{(k)} - \widehat{\mathbf{m}})' (\mathbf{v}\ \mathbf{X}) [(\mathbf{v}\ \mathbf{X})' \mathbf{W} (\mathbf{v}\ \mathbf{X})]^{-1} (\mathbf{v}\ \mathbf{X})' (\widehat{\mathbf{m}}^{(k)} - \widehat{\mathbf{m}}). \tag{A.8}$$

Similarly,

$$\sup_{\delta} Q^{(k)}(0, \delta) = (\widehat{\mathbf{m}}^{(k)} - \widehat{\mathbf{m}})' \mathbf{X} (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}' (\widehat{\mathbf{m}}^{(k)} - \widehat{\mathbf{m}}). \tag{A.9}$$

It follows from (2.9), (A.8), (A.9) and additional matrix manipulations that

$$\sup_{\alpha,\delta} Q^{(k)}(\alpha,\delta) - \sup_{\delta} Q^{(k)}(0,\delta) = \frac{|\mathbf{v}'\mathbf{U}(\mathbf{m}^{(k)} - \widehat{\mathbf{m}})|^2}{\mathbf{v}'\mathbf{V}\mathbf{v}},$$

and Theorem 1(b) follows from (A.7).

A.2 *Proof of Theorem* 3: *Logistic model*

Consider the model

$$m_i \sim \text{ Binomial}(p_i), \text{ where } \log\Big(\frac{p_i}{1-p_i}\Big) = \alpha v_i + \sum_{j=1}^{J} \beta_j x_{ij} \text{ for } 1 \le i \le u, \tag{A.10}$$

with $v_i = 1$ for $i \le a$ and $v_i = 0$ for $i > a$ for some $1 \le a \le u$. The log likelihood under (A.10) is

$$\ell(\alpha, \beta_1, \ldots, \beta_J) = \sum_{i=1}^{u} \Big\{ m_i\Big(\alpha v_i + \sum_{j=1}^{J} \beta_j x_{ij}\Big) - n_i \log\Big[1 + \exp\Big(\alpha v_i + \sum_{j=1}^{J} \beta_j x_{ij}\Big)\Big]\Big\} + \text{ constant.}$$

Let $\widehat{p}_i$ be the fitted-values at $(0, \widehat{\beta})$. Then (A.2) holds and in place of (A.3), the partial derivatives of ℓ at $(0, \widehat{\beta})$ satisfy

$$\begin{aligned} \ell_0 &= \sum_{i=1}^{a} (m_i - \widehat{m}_i), \quad \ell_{00} = -\sum_{i=1}^{a} n_i \widehat{p}_i(1-\widehat{p}_i), \\ \ell_{0j} &= -\sum_{i=1}^{a} x_{ij} n_i \widehat{p}_i(1-\widehat{p}_i), \quad \ell_{jh} = -\sum_{i=1}^{u} x_{ij} x_{ih} n_i \widehat{p}_i(1-\widehat{p}_i). \end{aligned}$$

Hence the second-order expansion of $2[\ell(\alpha, \widehat{\beta}+\delta) - \ell(0, \widehat{\beta})]$ is again (A.4), but with $\mathbf{W}$ a $u \times u$ diagonal matrix with entries $n_i\widehat{p}_i(1-\widehat{p}_i)$. Theorem 2(a) then follows from (A.5). For the logistic model, (A.6) again holds and Theorem 2(b) follows from (A.7)–(A.9).

References

[1] Kulldorff M, Information Management Services. *SaTScan User Guide*: http://www.satscan.org, 2006.

[2] Dwass M. Modified randomization tests for non-parametric hypotheses. *Annals of Mathematical Statistics* 1957; **28**:181–187.

[3] Kulldorff M. A spatial scan statistic. *Communications in Statistics–Theory and Methods* 1997; **26**:1481–1496.

[4] Cressie N, Chan NH. Spatial modeling of regional variables. *Journal of the American Statistical Association* 1989; **84**:393–401.

[5] Chan HP. Detection of spatial clustering with average likelihood ratio test statistics. *Annals of Statistics*, 2009; **37**:3985–4010.

[6] Pollard D. *Convergence of Stochastic Processes*. Springer: New York, 1984.

[7] Shorack G, Wellner J. *Empirical Processes with Applications to Statistics*. John Wiley: New York, 1986.

[8] Jung I. A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statistics in Medicine* 2009; **28**:1131–1143.

[9] McCullough P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1995.

[10] Loh J, Zhu Z. Accounting for spatial correlation in the scan statistic. *Annals of Applied Statistics* 2007; **2**:560–584.

[11] Zhang T, Lin G. Spatial scan statistics in loglinear models, *Computational Statistics & Data Analysis* 2009; **53**:2851-2858.

[12] Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, *American Journal of Public Health* 1998; **88**:1377–1380.

[13] Agresti A. *Categorical Data Analysis*, 2nd edn. John Wiley: New York, 2002.