Quasi-likelihood Scan Statistics for Detection of Spatial Clusters with Covariates

Pei-Sheng Lin

Division of Biostatistics and Bioinformatics, National Health Research Institutes & Department of Mathematics, National Chung Cheng University

17 December 2011
Outline

1 Background

2 Scan Statistics

3 Models and Selection Procedure

4 Simulation and Data Analysis
Motivation: A study of enterovirus pattern in Taiwan.

- Collected by CDC of Taiwan from 1999 to 2009.
- In 2003, 18224 cases reported in 89 administrative regions of the North Taiwan.
- Would temperature/humidity affect the disease rate?

How to identify geographic clusters incorporated with spatial correlation and environmental covariates?
Enterovirus Data

**Figur:** (a) Grey-scale map of the observed enterovirus rates. A region with a slash mark means missing value. (b) Locations of estimated clusters by the QL scan statistic.
Two main methodologies to find geographic clusters:

- Likelihood ratio tests for localized clusters (e.g. Chan, 2009, Annals of Statistics).
- Disease mapping on a hierarchical Bayesian model (e.g. Gangnon and Clayton, 2003, Statistics in Medicine).

However, current methods could neither straight incorporate ecological covariates nor account for spatial correlation (Zhang and Lin, 2009, Biometrics).

We explore a scan method based on quasi-likelihood functions for localized clusters.
Scan Statistics

- Basic concept: moving windows across space or time.
- Each distinct window induces a statistical model: difference in risk inside v.s. outside window
- Many potential choices for windows in space: circles, ellipses, arbitrary shapes
- We use circular windows for power consideration.
Potential Clusters

- We use centroid $s_k$ to denote region $k$.
- For each $s_k$, we sort the distances from $s_k$ to other centroids, say $r_k,(1) < r_k,(2) < \cdots < r_k,(m_k)$.
- Then $B\{s_k,r_k,(j)\}$ is a potential cluster, $j = 1, \ldots, m_k$.
- Each potential cluster contains at most 35% of the total population.
Latent Process

- $Y_i$: responses correlated through a latent Gaussian process $\epsilon$.
- $Y_i|\epsilon_i \sim \text{Poisson}(\theta_i^*)$, where

$$\theta_i^* = E_i \exp \left\{ x_i' \beta + \sum_{k=1}^{K} \xi_k \delta_{B_k}(s_i) + \epsilon_i \right\}.$$  \hspace{1cm} (1)

- $B_k$ and $\xi_k$ denote the $k$th cluster and associated coefficient, and $x_i$ and $\beta$ correspond to covariates.
- $(\epsilon_i, \epsilon_j) \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \rho_{i,j} \\ \sigma^2 \rho_{i,j} & \sigma^2 \end{pmatrix} \right\}$. 
Spatial Linear Mixed Model

By moment generating functions, (1) becomes

$$\theta_i = E(Y_i) = E_i \exp \left\{ 0.5\sigma^2 + x_i'\beta + \sum_{k=1}^{K} \xi_k \delta_{B_k}(s_i) \right\} \tag{2}$$

with a covariance structure given by

$$\text{var}(Y_i) = \theta_i + \theta_i^2 \{ \exp(\sigma^2) - 1 \},$$

$$\text{cov}(Y_i, Y_j) = \theta_i \theta_j \{ \exp(\sigma^2 \rho_{i,j}) - 1 \}. \tag{3}$$

In (2), $\xi_i$ could be positive or negative. So we can detect hot-spot and cool-spot clusters.
Two-stage Estimation

To estimate parameters, the first step is to find a working correlation model:

1. Obtain an initial estimate \( \{\hat{\beta}_0, \hat{\xi}_0\} \) by the traditional GLM method.
2. Let \( \hat{\epsilon}_i^{(0)} = \log(Y_i) - \log(\hat{\theta}_i^{(0)}) \), and obtain \( \hat{\sigma}^{(1)} \) and \( \hat{\rho}_{i,j}^{(1)} \) by variogram.
3. Plug \( \hat{\sigma} \) and \( \hat{\rho}_{i,j}^{(1)} \) to (3) to get a working covariance matrix.
**Quasi-likelihood Estimating Equation**

The second estimation step is to use a QL estimating equation

\[ Q(\hat{\beta}, \hat{\xi}; \hat{\sigma}, \hat{\rho}_{i,j}) = D'(\hat{\beta}, \hat{\xi}; \hat{\sigma})V^{-1}(\beta, \xi; \sigma, \rho_{i,j})\{Y - \theta(\beta, \xi; \sigma)\} = 0. \]

We derive a new estimate by a Newton-Raphson iteration

\[ \left\{ \hat{\beta}^{(k+1)}, \hat{\xi}^{(k+1)} \right\}' = \left\{ \hat{\beta}^{(k)}, \hat{\xi}^{(k)} \right\}' + \left[ (D'V^{-1}D)^{-1}D'V^{-1}\{Y - \theta(\beta, \xi; \sigma)\} \right] \bigg| \left\{ \hat{\beta}^{(k)}, \hat{\xi}^{(k)}, \hat{\sigma}, \hat{\rho}_{i,j} \right\}', \]
Asymptotic Properties

Assumption: (a) Number of misspecified regions to a cluster is $o(n)$. (b) Working correlation models $\rho_{i,j}^* = o(h^{-d})$.

Theorem: Under the assumptions, the QL estimates $(\hat{\beta}, \hat{\xi})$ are consistent. Furthermore, if $n^{-1/2}||\theta - \theta^*||_1 \rightarrow 0$ uniformly, then

$$n^{1/2}\{(\hat{\beta}, \hat{\xi}) - (\beta, \xi)\} \rightarrow N\{0, I_1^{-1}(\beta^*)I_0(\theta^*)I_1^{-1}(\theta^*)\}$$
Multiple Testing Problem

- Let $H$ denote the null hypothesis of no geographic cluster.
- When using the QL scan statistic, we would face a multiple testing problem.
- We adapt a parametric bootstrapping approach to find a correction $p$-value $p_0$. 
Parametric Bootstrapping

(a) We compute $E_i = n_i \hat{p}$, where $\hat{p} = \frac{\sum Y_i}{\sum n_i}$, and fit a null marginal model $\theta_i = E_i \exp(\beta_0 + \beta_1 x_i)$. Let $(\tilde{\beta}_0, \tilde{\beta}_1)$ and $(\tilde{\sigma}, \tilde{\rho}_{i,j})$ be the two-stage estimates.

(b) We simulate $Y^{(j)}$ from models (2) and (3) based on $(\tilde{\beta}_0, \tilde{\beta}_1)$ and $(\tilde{\sigma}, \tilde{\rho}_{i,j})$, $j = 1, \ldots, n_0$.

(c) For the $j$th simulated data, we use the QL scan statistic get p-values of all potential clusters.
Estimated Clusters

(d) Let $\hat{p}^{(j)}$ be the smallest $p$-value for the $j$th simulated data, $j = 1, \ldots, n_0$.

(e) We sort $\mathcal{P} = \{\hat{p}^{(1)}, \ldots, \hat{p}^{(n_0)}\}$ and set $p_0$ to be the 95% quantile of $\mathcal{P}$.

(f) For the real data, we compare $\hat{p}_{i,(j)}$ and $p_0$. A potential cluster $B\{s_i, r_{i,(j)}\}$ enters the final screening if $\hat{p}_{i,(j)}$ is less than $p_0$. 
Let $\mathcal{B} = \{\hat{B}_1^*, \ldots, \hat{B}_m^*\}$ be a set of estimated clusters from bootstrapping. Some of them would have overlapping regions.

For the overlapping clusters, we screen them with a quasi-deviance (Lin, 2010)

$$D(\hat{\xi}_i^*, \hat{\xi}_j^*) = 0.5\left\{ \theta(\hat{\xi}_i^*; \cdot) - \theta(\hat{\xi}_j^*; \cdot) \right\}^T$$

$$\times \left[ V^{-1}(\hat{\xi}_i^*; \cdot)\{Y - \theta(\hat{\xi}_i^*)\} + V^{-1}(\hat{\xi}_j^*; \cdot)\{Y - \theta(\hat{\xi}_j^*)\} \right].$$
Cluster Reformation

Using the quasi-deviance criteria to overlapping clusters, we

- discard \( \hat{B}_j^\ast \) if \( D(\hat{\xi}_i^\ast, \hat{\xi}_j^\ast) > \chi^2_{1,0.95} \), or,
- combine \( \hat{B}_i^\ast \) and \( \hat{B}_j^\ast \) if \( D(\hat{\xi}_i^\ast, \hat{\xi}_j^\ast) \leq \chi^2_{1,0.95} \), and
- sequentially screen clusters from the one with a smallest p-value to the one with a largest p-value.

For the final \( K \) disjoint clusters \( \hat{B}_1, \ldots, \hat{B}_K \), the two-stage estimation procedure gives

\[
\hat{\theta}_i = E_i \exp \left\{ 0.5\hat{\sigma}^2 + \mathbf{x}'_i \hat{\beta} + \sum_{k=1}^{K} \hat{\xi}_k \delta_{\hat{B}_k}(\mathbf{s}_i) \right\}.
\]
Simulation Scenario:

- $\epsilon$: GRF with mean 0, $\sigma^2 = 0.25$ and $\rho = 0.7$.
- $Y_i | \epsilon_i$: Poisson with intensity
  $\theta_i^* = \exp\{\beta_0 + \xi_1 \delta_{B_1}(s_i) + \epsilon_i\}$.
- parameters: $\beta_0 = -0.6$, $\xi = 0, 0.5, 0.75, 1.0, \text{ or } 1.5$.
- cluster: $B_1$ is a cluster centered at Zhong-li city with 15 adjacent regions.
Kulldorff’s LR Test

When no covariates, Kulldorff (1997) suggested to take a maximum value of likelihood ratio over all possible clusters. That is,

$$LR_{max} = \max_{k,j} LR_{k,r_k,(j)},$$

where

$$LR_{k,r_k,(j)} = \sup \mathcal{L}(\Theta_1 | y) / \sup \mathcal{L}(\Theta_0 | y).$$

We compare the QL scan statistic with LR.
Comparison of Power Functions

Figure: A comparison of power based on 1000 simulations. (a) The pre-specified cluster. (b) The estimated power functions.
Tabel: Conditional sample means and standard errors for QL and GLM scan statistics.

<table>
<thead>
<tr>
<th></th>
<th>QL</th>
<th>GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_1$</td>
<td>$\hat{\beta}_2$</td>
</tr>
<tr>
<td>$(\beta^*_1, \beta_2, \xi_1, \xi_2) = (-0.475, 0.5, 0.5, 0.5)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>-0.53</td>
<td>0.48</td>
</tr>
<tr>
<td>Asym SE</td>
<td>(0.189)</td>
<td>(0.266)</td>
</tr>
<tr>
<td>Sample SE</td>
<td>(0.158)</td>
<td>(0.254)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(\beta^*_1, \beta_2, \xi_1, \xi_2) = (-0.475, 0.5, 1.0, 1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.45</td>
</tr>
<tr>
<td>Asym SE</td>
<td>(0.195)</td>
</tr>
<tr>
<td>Sample SE</td>
<td>(0.156)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(\beta^*_1, \beta_2, \xi_1, \xi_2) = (-0.475, 0.5, 1.5, 1.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.44</td>
</tr>
<tr>
<td>Asym SE</td>
<td>(0.206)</td>
</tr>
<tr>
<td>Sample SE</td>
<td>(0.196)</td>
</tr>
</tbody>
</table>
Data Analysis

2003 North Taiwan Data:

- 21 regions without observations.
- Environmental covariates temperature and humidity were standardized to be in (-1,1).

We use the QL scan statistic for the 2003 data to get

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.01</td>
<td>1.62</td>
<td>0.95</td>
</tr>
<tr>
<td>Temp</td>
<td>14.24</td>
<td>4.16</td>
<td>0.00</td>
</tr>
<tr>
<td>Humidity</td>
<td>16.61</td>
<td>4.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>-0.75</td>
<td>0.41</td>
<td>0.07</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.93</td>
<td>0.34</td>
<td>0.01</td>
</tr>
</tbody>
</table>
The estimated model for the disease rate of the $i$th region is

$$
\hat{\theta}_i = 0.0032 n_i \exp\{14.24 (\text{temp})_i + 16.61 (\text{humidity}_i) - 0.75 \delta_{B_1}(s_i) + 0.93 \delta_{B_2}(s_i)\},
$$

with a correlation estimate $\hat{\rho}_{i,j} = 0.94 \exp(-\|s_i - s_j\|/3.32)$ and $\hat{\sigma}^2 = 0.45$. 
**Figur:** (a) Grey-scale map of the observed enterovirus rates. A region with a slash mark means missing value. (b) Locations of estimated clusters by the QL scan statistic.
Future Work

- Variogram may not work for some potential cluster models.
- Non-stationary correlation models for latent processes.
- Add temporal random effects for a Bayesian dynamic linear model.