Local Likelihood Models for Disease Cluster Modeling: A Space-Time Extension

1. Introduction

Data-dependent clustering (Lawson, 2006; Hossain and Lawson, 2005) assumed that the relative risk of a disease is defined to be a function of the 'local' concentration of cases. This is the idea of local likelihood and the model based on local likelihood has found a novel application to disease cluster modeling. Thus, the local relative risk estimator is based simply on the relation of a data item to other data within a lasso distance. The optimum values for lasso distances are estimated based on the closeness of the observed cases. In the process of estimation, the data are aggregated based on the lasso distances and it is assumed that the aggregated data are independent of each other for a given lasso distance. Because of this way of aggregating observed data, the local likelihood model that we have proposed for spatial data in Lawson (2006) and Hossain and Lawson (2005) has found good application in disease cluster modeling. This paper extends the spatial local likelihood model to spatio-temporal settings.

In our earlier work for spatial data, we assume a circular window for each data point (e.g., the point can be centroid of a county for count data or a residential location for event data) whose radius is a lasso parameter. The extension that we propose for space-time data is based on extending the circular window for the lasso parameter to a cylindrical window where the base representing space and height representing time. The optimum values for the radius of base and the length of height are determined by using the spatial and temporal associations that exist in the data. As for the estimation procedure, we employ a Bayesian hierarchical modeling approach with a joint implementation of Gibbs and Metropolis-Hasting MCMC computational methods to obtain posterior estimates of all model parameters.

To judge the relative performances of this model, we made a comparison with the models proposed as space-time extension of random effect models (Knorr-Held and Besag, 1998; Besag, York and Mollie, 1991). In order to check the ability of each models in recovering clusters, we extended the criteria those proposed in Hossain and Lawson (2006) to space-time setting. These criteria are proposed to recover spatial clusters. Here, we extended them to recover spatio-temporal clusters. To define the spatial and spatio-temporal clustering, we follow the definition
provided by Clark and Lawson (2002). These authors defined three types of clustering; spatial, temporal and spatio-temporal in order to explain the complexities involved in cluster modeling for space-time data. The elevated risks if it persists to a specific region throughout the whole study period then it is a spatial cluster, if it persists throughout the whole study region but for a limited period of time then it is a temporal cluster, and it is a spatio-temporal cluster if it persists to a specific region for a limited period of time. We adopt these definitions in our simulation experiment where the data are generated to have a spatial cluster and two spatio-temporal clusters.

The remaining sections are organized as follows: in section 2 we illustrate the local likelihood model for space-time data. Section 3 introduces the Ohio lung cancer mortality data and the simulation experiment.

2. Model Development

Let us assume that the data is observed as \( O_i \) to denote the observed count of death and \( E_i \) to denote the expected count for county \( i \) and year \( t \) where \( i = 1, \ldots, I \) and \( t = 1, \ldots, T \). The expected counts are case-mix adjusted and the adjustment is done for important confounders. It is a common practice to assume a Poisson model for count data where the events are rare. So, at the first level of hierarchy, we model the observed count as

\[
O_i \mid \delta_i, \theta_i \sim \text{Poisson} \left( E_i, \theta_i \right) \tag{2.1}
\]

where \( O_i = \sum_{d_i^{(x)} \leq \kappa_i, d_i^{(y)} \leq \gamma_i} O_{tt} \), \( E_i = \sum_{d_i^{(x)} \leq \kappa_i, d_i^{(y)} \leq \gamma_i} E_{tt} \), \( i = 1, \ldots, I \) and \( t = 1, \ldots, T \). The space-time lasso can be defined as \( \delta_i = \{ e^{\kappa_i}, e^{\gamma_i} \} \). The Euclidean distance between the centroids of counties \( i \) and \( i' \) is denoted by \( d_i^{(x)} \) and between years \( t \) and \( t' \) is denoted by \( d_i^{(y)} \). The spatial and temporal distances are standardized to have the values between \( \left[ 0, \sqrt{2} \right] \) and \( [0, 1] \), respectively. One of the simplest way of defining these lasso parameters, \( \kappa_i \) and \( \gamma_i \), as spatially and temporally uncorrelated random effects. So, the spatial lasso parameter, \( e^{\kappa_i} \), is the radius and the temporal lasso parameter, \( e^{\gamma_i} \), is the half-height of \((i, t)\)-th cylinder.
At the second level of hierarchy, the logarithm of relative risk, $\theta_i$, is defined as

$$\nu_i = \log (\theta_i) = \rho + \kappa_i + \gamma_i + \epsilon_i + \xi_i + \lambda_{it}$$  \hspace{1cm} (2.2)

where $\rho$ is an overall mean effect on log-relative risk, $\kappa_i$ and $\gamma_i$ are two unstructured spatial and temporal random effects (as mentioned above), $\epsilon_i$ and $\xi_i$ are two structured spatial and temporal random effects and $\lambda_{it}$ is a space-time interaction random effects. The space-time interaction effects can have various forms depending on structured/unstructured spatial and structured/unstructured temporal interaction. In this paper, for simplicity, we have considered the interaction of unstructured spatial and unstructured temporal effects. A more complicated form for interaction effect, e.g. interaction of structured spatial and structured temporal effects can be considered as similar as Knorr-Held (2000). In practice, choice of a simpler or a more complicated form of interaction effect can be justified by model goodness-of-fit criteria, e.g., DIC (Spiegelhalter, Best, Carlin and Linde, 2002).

The priors are:

- $\rho \sim \mathcal{N}(0, k_{\rho})$, $k_{\rho}$ is assumed to be known,
- $\kappa_i | \sigma^2 \sim \mathcal{N}(\mu_\kappa, \sigma^2_\kappa)$, $\mu_\kappa$ is assumed to be known,
- $\gamma_i | \sigma^2 \sim \mathcal{N}(\mu_\gamma, \sigma^2_\gamma)$, $\mu_\gamma$ is assumed to be known,
- $\epsilon_i | \{\epsilon_i \}, \sigma_\epsilon \sim \mathcal{N}\left(\overline{\epsilon}_i, \frac{\sigma_\epsilon}{d_i}\right)$,
- $\xi_i | \xi_{i-1}, \sigma^2 \sim \mathcal{N}(\xi_{i-1}, \sigma^2)$, and,
- $\lambda_{it} | \sigma^2 \sim \mathcal{N}(0, \sigma^2_{\lambda})$.

The choices of hyperpriors for all variance parameters are assumed to have inverse gamma distribution as $\text{Inv-\gamma}(a,b)$. The scale and shape parameters for inverse gamma are set to 2.01 and 1.01 to reflect our prior belief about variance parameters are equivalent to 1 prior measurement and having common variance 100. The variance parameter for prior $\rho$, $k_{\rho}$ is set to a large number, 10,000. The mean parameters for priors $\kappa$ and $\gamma$ are set to a negative value,
The justification of setting the means to $-20.0$ is that a priori we are assuming the observed data contains no cluster.

3. Ohio Lung Cancer Data

We use the lung cancer mortality in 88 counties of Ohio State for the year 1968-88 to illustrate the method. The available variables are observed and expected death counts for $i$-th county and $t$-th year where $i = 1, \ldots, 88$ and $t = 1, \ldots, 21$. The case-mix adjusted expected deaths are computed for gender and race case mixing. The observed SMR (figure 1), the ratio of observed to expected counts, shows no spatial or spatio-temporal clustering. For this reason, we conduct a simulation experiment for the given Ohio geography where the true risks are assigned in a way to have spatial and/or spatio-temporal clusters. The simulated data also makes feasible a relative comparison between space-time local likelihood model as described above and space-time extension of BYM model, where the true risks are known.

The true risks, $\{\theta_{ij}^{true}\}$, are generated randomly from Uniform$(0.5, 1.5)$. One spatial cluster of elevated risk is introduced by adding excess risk to Carroll county in the west of Ohio. In a similar way, two spatio-temporal clusters of elevated risks is introduced: one in the north and constitutes of Ashland, Wayne, Richland, Medina, Lorain, Knox, Huron and Holmes counties for the years 1977 and 1978, and the other one near the middle and constitutes of Coshocton, Tuscarawas, Muskingum, Licking, Knox, Holmes and Guernsey counties for the years 1983 and 1984. Thus, the true risk scenarios include a range of clustering types; spatial clustering and spatio-temporal clustering within the study region. Figure 2 shows the thematic map or true relative risks. Given the true relative risks define under the model and the expected counts from the real data, we conditionally simulated 30 sets of observed counts for each county and year from a Poisson distribution with:

$$y_{its} | e_{it}, \theta_{it}^{true} \sim \text{Poisson}(e_{it} \theta_{it}^{true}); \quad i = 1, \ldots C, \quad t = 1, \ldots, T \quad \text{and} \quad s = 1, \ldots, 30$$

where $\theta_{ij}^{true}$ is assumed to be the true value for $\theta_{ij}$. The number of datasets is fixed to 30 only to reduce the computational time.

Results: Results will be presented in the talk.


