Modeling Areal Data: Intro

Spatial Smoothing

Areal Data Models: Disease Mapping

Areal data with counts is often associated with disease mapping, where there are two quantities for each areal unit:

One way to think about the expected counts is

However note that $\bar{r}$, and hence, E_i is a not fixed, but is a function of the data.

An alternative is to use some standard rate for a given age group, such that $E_i = \sum_j n_{ij} r_j$.

Traditional Models

Often counts are assumed to follow the Poisson model where

Then the MLE of η_i is $\frac{Y_i}{E_i}$.

Poisson-Gamma Model

Consider the following framework

For the Poisson sampling model, the gamma prior is conjugate. This means that the posterior distribution $p(\eta_i|y_i)$ is also a gamma distribution, and in particular

The mean of this distribution is

$$\begin{aligned} E(\eta_i|\boldsymbol{y}) = E(\eta_i|y_i) &= \frac{y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}} \\ &= \frac{E_i(\frac{y_i}{E_i})}{E_i + \frac{\mu}{\sigma^2}} + \frac{(\frac{\mu}{\sigma^2})\mu}{E_i + \frac{\mu}{\sigma 2}} \\ &= w_i SMR_i + (1 - w_i)\mu, \end{aligned}$$

Poisson-lognormal models

The model can be written as

$$\begin{aligned} Y_i|\psi_i &\sim Poisson(E_i \exp(\psi_i)) \\ \psi_i &= \boldsymbol{x_i^T}\boldsymbol{\beta} + \theta_i + \phi_i \end{aligned}$$

Brook's Lemma and Markov Random Fields

To consider areal data from a model-based perspective, it is necessary to obtain the joint distribution of the responses

$$p(y_1, \ldots, y_n).$$

From the joint distribution, the *full conditional distribution*

$$p(y_i|y_j, j \neq i),$$

is uniquely determined.

When the areal data set is large, working with the full conditional distributions can be preferred to the full joint distribution.

More specifically, the response Y_i should only directly depend on the neighbors,

Markov Random Field

The idea of using the local specification for determining the global form of the distribution is

An essential element of a MRF is a *clique*, which is a group of units where each unit is a neighbor of all units in the clique

A *potential function* is a function that is exchangeable in the arguments. With continuous data a common potential is $(Y_i - Y_j)^2$ if $i \sim j$ (i is a neighbor of j).

Gibbs Distribution

A joint distribution $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of Y_i only through the potential on cliques.

Mathematically, this can be expressed as:

$$p(y_1, \ldots, y_n) \propto \exp\left(\gamma \sum_k \sum_{\alpha \in \mathcal{M}_k} \phi^{(k)}(y_{\alpha_1}, y_{\alpha_2}, \ldots, y_{\alpha_k})\right),$$

where $\phi^{(k)}$ is a potential of order k, $\mathcal{M}_k$ is the collection of all subsets of size $k = 1, 2, ...$(typically restricted to 2 in spatial settings), α indexes the set in $\mathcal{M}_k$.

Hammersley-Clifford Theorem

The Hammersley-Clifford Theorem demonstrates that if we have a MRF that defines a unique joint distribution,

The converse was later proved, showing that a MRF could be sampled from the associated Gibbs distribution.

Model Specification

With continuous data, a common choice for the joint distribution is the pairwise difference

$$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2}\sum_{i,j}(y_i - y_j)^2 I(i \sim j)\right)$$

Then the full conditional distributions