LECTURE 19: AREAL DATA MODELING

CLASS INTRO

INTRO QUESTIONS

With Areal Data we have discussed:

1. Data Visualization
2. Proximity Matrices
3. Measures of Spatial Association
4. Spatial Smoothing
5. Technical Details for Autocorrelated Models

AREAL DATA OVERVIEW

Today

- Disease Models
- CAR / SAR Models

Next

- Model fitting with `spdep` and `JAGS`

AREAL DATA MODELS

DISEASE MAPPING

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

$$Y_i = \text{ observed number of cases of disease in county i}$$
$$E_i = \text{ expected number of cases of disease in county i}$$

EXPECTED COUNTS

- One way to think about the expected counts is
  $E_i = n_i \bar{r} = n_i \left(\frac{\sum_i y_i}{\sum_i n_i} \right)$,
  where $\bar{r}$ is the overall disease rate and n_i is the population for region i.
- However note that $\bar{r}$, and hence, E_i is a not fixed, but is a function of the data. This is called *internal standardization*.
- An alternative is to use some standard rate for a given age group ($r_{j\cdot}$), such that $E_i = \sum_j n_{ij} r_{j\cdot}$. This is *external standardization*.

TRADITIONAL MODELS

- Often counts are assumed to follow the Poisson model where
  $Y_i|\eta_i \sim Poisson(E_i\eta_i)$,
  where η_i is the relative risk of the disease in region i.
- **Discuss:** how does η_i impact our expectation about the number of diseases cases?
- **Discuss:** How can η_i be interpreted?

TRADITIONAL MODELS

- Often counts are assumed to follow the Poisson model where
  $Y_i|\eta_i \sim Poisson(E_i\eta_i)$,
  where η_i is the relative risk of the disease in region i.
- Then the MLE of η_i is $\frac{Y_i}{E_i}$. This quantity is known as the *standardized morbidity ratio* (SMR).
- Note the SMR is multiplicative rather than additive.

POISSON-GAMMA MODEL

- Consider the following framework
  $Y_i|\eta_i \sim Po(E_i\eta_i)$, i = 1, ..., I
  The goal is estimate η_i and eventually include spatial effects.

- We start with a prior for η_i as
  $\eta_i \sim Gamma(a, b)$,
  where the gamma distribution has mean $\frac{a}{b}$ and variance is $\frac{a}{b^2}$.

- This can be reparameterized such that $a = \frac{\mu^2}{\sigma^2}$ and $b = \frac{\mu}{\sigma^2}$.

POISSON-GAMMA CONJUGACY

- 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.
- In particular the posterior distribution $p(\eta_i|y_i)$ is $Gamma(y_i + a, E_i + b)$.

BAYESIAN POINT ESTIMATE

- The mean of this distribution is

$$E(\eta_i|\boldsymbol{y}) = E(\eta_i|y_i) = \frac{y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}}$$

$$= \frac{E_i(\frac{y_i}{E_i})}{E_i + \frac{\mu}{\sigma^2}} + \frac{(\frac{\mu}{\sigma^2})\mu}{E_i + \frac{\mu}{\sigma 2}}$$

$$= w_i SMR_i + (1 - w_i)\mu,$$

where $w_i = \frac{E_i}{E_i + (\mu/\sigma^2)}$

- **Discuss:** this result and suggestions for selecting μ and σ^2.

POISSON-LOGNORMAL MODELS

- Unfortunately the Poisson-gamma framework does not easily permit spatial structure with the η_i and a univariate gamma distribution.
- A relatively new option is to use the **multivariate-log gamma distribution** as the prior.
- A common alternative is to use the Poisson-lognormal model.

POISSON-LOGNORMAL MODELS

The model can be written as

$$Y_i|\psi_i \sim Poisson(E_i \exp(\psi_i))$$

$$\psi_i = \boldsymbol{x_i^T} \boldsymbol{\beta} + \theta_i + \phi_i$$

where $\boldsymbol{x_i}$ are spatial covariates, θ_i corresponds to region wide heterogeneity, and ψ_i captures local clustering.

DATA SIMULATION EXERCISE

1. Simulate and visualize data following the Poisson-gamma framework.
2. Now think about adding spatial structure using the log-normal structure. How would you generate θ_i and ϕ_i?

CONDITIONAL AUTOREGRESSIVE MODELS

GAUSSIAN MODEL

- Suppose the full conditionals are specifed as
  $$Y_i | y_j, j \neq i \sim N\left(\sum_j b_{ij} y_j, \tau_i^2\right)$$
- Then using Brooks' Lemma, the joint distribution is
  $$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2}\boldsymbol{y}^T D^{-1}(I - B)\boldsymbol{y}\right),$$
  where B is a matrix with entries b_{ij} and D is a diagonal matrix with diagonal elements $D_{ii} = \tau_i^2$.

GAUSSIAN MODEL

- The previous equation suggests a multivariate normal distribution, but $D^{-1}(I - B)$ should be symmetric.
- Symmetry requires
$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}, \quad \forall \quad i, j$$
- In general, B is not symmetric, but setting $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$ satisfies the symmetry assumptions (given that we assume W is symmetric)

GAUSSIAN MODEL

- Now the full conditional distribution can be written as
  $$Y_i | y_j, j \neq i \sim N\left(\sum_j w_{ij} y_j / w_{i+}, \tau^2 / w_{i+}\right)$$
- Similarly the joint distribution is now
  $$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2} \boldsymbol{y}^T (D_w - W) \boldsymbol{y}\right)$$
  where D_w is a diagonal matrix with diagonal entries $(D_w)_{ii} = w_{i+}$

GAUSSIAN MODEL

- The joint distribution can also be re-written as
  $$p(y_1, \dots, y_n) \propto \exp\left(-\frac{1}{2\tau^2} \sum_{i \neq j} w_{ij}(y_i - y_j)^2 \right)$$
- However, both these formulations results in an improper distribution. This could be solved with a constraint, such as $Y_i = 0$.
- The result is the joint distribution is improper, despite proper full conditional distributions. This model specification is often referred to as an *intrinsically autoregressive* model (IAR).

IAR

- The IAR cannot be used to model data directly, rather this is used a prior specification and attached to random effects specified at the second stage of the hierarchical model.
- The impropriety can be remedied by defining a parameter ρ such that $(D_w - W)$ becomes $(D_w - \rho W)$ such that this matrix is nonsingular.
- The parameter ρ can be considered an extra parameter in the CAR model.

POSTERIOR DISTRIBUTION

- With or without ρ, $p(\boldsymbol{y})$ (or the Bayesian posterior when the CAR specification is placed on the spatial random effects) is proper.
- When using ρ, the full conditional becomes
$$Y_i | y_j, j \neq i \sim N\left(\rho \sum_j w_{ij} y_j / w_{i+}, \tau^2 / w_{i+}\right)$$
- The authors state, “we do not take a position with regard to propriety or impropriety in employing CAR specifications”

SIMULTANEOUS AUTOREGRESSION MODEL

SIMULTANEOUS AUTOREGRESSION MODEL

- Rather than specifying the distribution on $\boldsymbol{Y}$, as in the CAR specification, the distribution can be specified for $\boldsymbol{\epsilon}$ which induces a distribution for $\boldsymbol{Y}$.
- Let $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tilde{D})$, where $\tilde{D}$ is a diagonal matrix with elements $(\tilde{D})_{ii} = \sigma_i^2$.
- Now $Y_i = \sum_j b_{ij} Y_j + \epsilon_i$ or equivalently $(I - B)\boldsymbol{Y} = \boldsymbol{\epsilon}$.

SAR MODEL

- If the matrix $(I - B)$ is full rank, then $\boldsymbol{Y} \sim N\left(\boldsymbol{0}, (I - B)^{-1}\tilde{D}((I - B)^{-1})^T\right)$
- If $\tilde{D} = \sigma^2 I$, then $\boldsymbol{Y} \sim N\left(\boldsymbol{0}, \sigma^2\left[(I - B)(I - B)^T\right]^{-1}\right)$

CHOOSING B

There are two common approaches for choosing B

1. $B = \rho W$, where W is a contiguity matrix with entries 1 and 0. The parameter ρ is called the spatial autoregression parameter.
2. $B = \alpha \tilde{W}$ where $(\tilde{W})_{ij} = w_{ij}/w_{i+}$. The α parameter is called the spatial autocorrelation parameter.

SAR MODEL FOR REGRESSION

- SAR Models are often introduced in a regression context, where the residuals ($\boldsymbol{U}$) follow a SAR model.
- Let $\boldsymbol{U} = \boldsymbol{Y} - X\boldsymbol{\beta}$ and then $\boldsymbol{U} = B\boldsymbol{U} + \boldsymbol{\epsilon}$ which results in $\boldsymbol{Y} = B\boldsymbol{Y} + (I - B)X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Hence the model contains a spatial weighting of neighbors ($B\boldsymbol{Y}$) and a regression component ($(I - B)X\boldsymbol{\beta}$).
- What is the result of extreme cases for $B = 0$ or $B = I$.

OTHER NOTES

- The SAR specification is not typically used in a GLM setting.
- SAR models are well suited for maximum likelihood
- Without specifying a hierarchical form, Bayesian sampling of random effects is more difficult than the CAR specification.