Spatial generalized linear models

Applied Spatial Statistics
Non-Gaussian spatial data

► Thus far we have assumed the response $Y_i$ is Gaussian

► Often you can transform the data to be approximately Gaussian, e.g., define the response as $\log(Y_i)$

► Slight deviation from normality is fine, but what if the response is binary or a count?

► Assuming normality is clearly inappropriate and we need new methods
Motivating examples

- **Binary example:** $Y_i = 1$ if a species is observed at $s_i$ and $Y_i = 0$ otherwise

- **Count example:** $Y_i \in \{0, 1, 2, \ldots\}$ is the number of days below freezing at $s_i$ in the year 2000

- **Classification example:** $Y_i = 1$ if $s_i$ is a forest, $Y_i = 2$ it’s a desert, $Y_i = 3$ if it’s a city

- **Extreme example:** $Y_i$ is the maximum one-hour precipitation at $s_i$ in 2020
Review of the Gaussian spatial model

The standard model is model is \( Y_i = \mu_i + Z_i + \varepsilon_i \)

- The mean is the same as linear regression
  \[ \mu_i = \beta_0 + X_i \beta_1 + \ldots + X_{ip} \beta_p \]

- There are two error terms:
  - \( Z_i \) is spatially-correlated
  - \( \varepsilon_i \sim \text{Normal}(0, \tau^2) \) are independent across \( i \)

- Example: \( \mathbb{E}(Z_i) = 0 \) and \( \text{Cov}(Z_i, Z_j) = \sigma^2 \exp(-d_{ij}/\phi) \)
Review of the Gaussian spatial model

- The joint distribution of all $n$ observations is

$$Y \sim \text{Normal}(\mathbf{X}\beta, \Sigma(\theta))$$

where $\beta = (\beta_0, ..., \beta_p)$ and $\theta = (\sigma^2, \tau^2, \phi)$

- The likelihood as a function of $\beta$ and $\theta$

- This *marginalizes out* the $Z_i$ which requires taking a complicated integral

- This trick avoids estimating the $Z_i$, but does not work for most non-Gaussian models
Review of logistic regression

- Logistic regression is the most common analysis method for a binary response, $Y_i \in \{0, 1\}$

- Denote the mean as $E(Y_i) = \text{Prob}(Y_i = 1) = \pi_i$

- Thus $\text{Prob}(Y_i = 0) = 1 - \pi_i$

- We want to relate the mean and the linear predictor

$$\eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j \in (-\infty, \infty)$$

- Setting $\pi_i = \eta_i$ is wrong because $\pi_i$ must be between zero and one
Review of logistic regression

- We insert the inverse logistic function to ensure the mean is between zero and one,

\[ \pi_i = \expit(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \]

- This is equivalent to

\[ \logit(\pi_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j \]

where \( \logit(\pi) = \log\{\pi/(1 - \pi)\} \) is the log odds

- Interpretation: \( \beta_j \) is the increase in the log odds of \( Y_i = 1 \) if \( X_{ij} \) increases by one and all other covariates are held fixed
Review of Poisson regression

- Poisson regression is the most common analysis method for a count response, \( Y_i \in \{0, 1, 2, \ldots\} \)

- Often the count is associated with a known sampling effort variable \( N_i \), i.e., hours of effort or population size

- Denote the mean as \( E(Y_i) = N_i \lambda_i \) so \( \lambda_i \) is the expected count per unit effort

- We want to relate the mean and the linear predictor
  \[
  \eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j
  \]

- Setting \( \lambda_i = \eta_i \) is wrong because \( \lambda_i \) must be positive
Review of Poisson regression

- To ensure $\lambda_i$ is positive we set $\lambda_i = \exp(\eta_i)$

- This is equivalent to

  $$\log(\lambda_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j$$

- Interpretation: The log of the mean increase by $\beta_j$ if $X_{ij}$ increases by one and all other covariates are held fixed

- Interpretation: The mean is multiplied by $\exp(\beta_j)$ if $X_{ij}$ increases by one and all other covariates are held fixed
Review of generalized linear models (GLMs)

- The response $Y_i$ can have any distribution: Gaussian, binomial, Poisson, Gamma, Negative binomial, etc.

- Whatever the distribution, define the mean as $\text{E}(Y_i) = \mu_i$

- The **link function** $g$ relates the mean and linear predictor,

  $$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j$$

- You can choose any link function that ensures that $\mu_i$ is in the appropriate range for any $X_i$ and $\beta$. 

Spatial GLMs

- A spatial GLM adds a spatial term to the linear predictor
  \[ \eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j + Z_i \]

- \( Z \) is a spatial process as in the Gaussian spatial model

- For example, \( \text{E}(Z_i) = 0 \) and \( \text{Cov}(Z_i, Z_j) = \sigma^2 \exp(-d_{ij}/\phi) \)

- Observations are assumed to be independent given the spatial random effects, \( Z_i \)

- A nugget is not included in \( Z_i \)
Spatial logistic regression

- Assume $Y_i \mid \pi_i \sim \text{Bernoulli}(\pi_i)$, independent over $i$ \(^1\)

- The probability $\text{Prob}(Y_i = 1) = \pi_i$ is modeled as

$$\text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + Z_i$$

- The $\beta_j$ are interpreted just like non-spatial logistic regression

\(^1\)A Bernoulli($\pi$) random variable is a Binomial($1, \pi$) random variable

\(^2\)If $Y$ is the number of successes in $n$ independent trials, each with success probability $\pi$, then $Y \sim \text{Binomial}(n, \pi)$
Random draw for $Z_1, \ldots, Z_n$
Plot of $\pi_i = \text{expit}(Z_i)$
Realization of $Y_i | \pi_i \sim \text{Bernoulli}(\pi_i)$
Spatial Poisson regression

- Assume $Y_i|\lambda_i \sim \text{Possion}(N_i\lambda_i)$, independent over $i$.

- $N_i$ is the known “offset term”

- The relative risk $\lambda_i$ is modeled as

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + Z_i$$

- The $\beta_j$ are interpreted just like non-spatial Poisson regression

---

An equivalent model used in some packages is $Y_i|\lambda_i \sim \text{Possion}(\lambda_i)$ where $\log(\lambda_i) = \log(N_i) + \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + Z_i$
Random draw for $Z_1, \ldots, Z_n$
Plot of $\lambda_i = \exp\{Z_i\}$
Realization of $Y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$
Spatial Gaussian regression

- The usual Gaussian model is a special case of a GLM

- Assume $Y_i | \eta_i \sim \text{Normal}(\eta_i, \tau^2)$, independent over $i$

- The mean $\eta_i$ is modeled as

$$\eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij} \beta_j + Z_i$$

- The link function is the identify function, $g(\eta) = \eta$
Spatial GLMs

- A spatial GLM assumes the responses are conditionally independent given $Z_i$
- The spatial terms $Z_i$ account for spatial dependence
- Even if $Z$ has a simple correlation structure, the marginal (over $Z$) correlation of $Y$ is hard to compute
- For example, in the logistic case we would need to be able to compute intractable quantities like

$$\text{Cov}\{\expit(Z_i), \expit(Z_j)\}$$
As mentioned in the introduction, it is hard to compute the joint likelihood.

For example, in the binary case, \( \text{Prob}(Y_i = Y_j = 1) \)

This makes MLE tricky.

However, a Bayesian analysis with MCMC is actually straightforward, but slow.

We’ll use \texttt{spBayes}, but there are other packages like \texttt{OpenBUGS}, \texttt{JAGS}, \texttt{INLA}, \texttt{STAN}, etc.