# STAT 8025

## Lecture 7: Non-Gaussian Spatial Data

Dr. Emily Lei Kang

Division of Statistics & Data Science
Department of Mathematical Sciences
University of Cincinnati



University of Cincinnati

# Introduction

- We can generalize our models to handle non-Gaussian data
  - Spatial binary data: $Y(\mathrm{s}) = 1$ if site s is disease and $Y(\mathrm{s}) = 0$ otherwise
  - Spatial count data: $Y(\mathrm{s}) \in \{0, 1, 2, \cdots\}$ is the number of days at site s with extreme temperature
- The multivariate normal distribution or GP is convenient, but this doesn't provide a good model for such non-Gaussian data (at least not directly)
- What models do you know for non-Gaussian data?

# Generalized Linear Model

- For some link function $g(\cdot)$, we have

$$\eta_i = g(E(Y_i)) = \mathrm{X}_i'\boldsymbol{\beta}$$

  - For binary data and logistic regression we have $g(r) = log\frac{r}{1-r}$
  - For count data and Poisson regression we have $g(r) = log(r)$
- Estimation is done via iterative reweighted least squares (IRWLS)
  - Consider the effective response $z \equiv \eta(x) + (Y - r(x)g'(r(x))$
  - Consider $var(z)$ and then perform weighted least squares (iteratively)

How to do this with spatial data?

- Peter Diggle: Spatial dependence is often induced by adding Gaussian random effects to a generalized linear model.
- Let's think about the Gaussian case:
  - We assume
    $$Y(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s}) + \epsilon(\mathrm{s})$$
    where $w(\cdot)$ is a GP with mean zero and covariance function $C(\cdot,\cdot)$; $\epsilon(\cdot)$ is iid Gaussian white noise with variance $\tau^2$
  - One way to write the model is:
    $$Y(\mathrm{s})|w(\mathrm{s}) \overset{ind}{\sim} N(\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s}), \tau^2)$$
    $$w(\cdot) \sim \mathcal{GP}(0, C(\cdot,\cdot))$$
  - Conditioning on $w(\cdot)$, $Y(\cdot)$ becomes spatially independent.
  - Marginally, (integrating out the random effects $w(\cdot)$), $Y(\cdot)$ is a GP with mean $\mu(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta}$ and covariance function $C(\cdot,\cdot) + \tau^2 1(\mathrm{s} = \mathrm{t})$.

# Spatial GLM

- We assume
$$w(\cdot) \sim \mathcal{GP}$$
with mean zero and covariance function $C(\cdot, \cdot)$.
- Given $w(\cdot)$, $Y(\mathrm{s})$ are independent with
$$g(E(Y(\mathrm{s}))) = \mathrm{X}(\mathrm{s})'\beta + w(\mathrm{s})$$
- Marginalizing over the random effects $w(\cdot)$, we will have spatially dependent $Y(\mathrm{s})$'s.
- However, it is often hard to get the closed-form expression for the marginal distribution of $Y(\cdot)$.

# Spatial GLM for binary data

Given $w(\mathrm{s})$,

$$Y(\mathrm{s}) \overset{indep.}{\sim} Bern(p(\mathrm{s}))$$

where

$$logit(p(\mathrm{s})) = \mathrm{X}(\mathrm{s})'\beta + w(\mathrm{s})$$

# Spatial GLM for count data

Given $w(\mathrm{s})$,

$$Y(\mathrm{s}) \overset{indep.}{\sim} Pois(\lambda(\mathrm{s}))$$

where

$$log(\lambda(\mathrm{s})) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s})$$

UNIVERSITY OF Cincinnati

Marginal distribution of $Y(\mathrm{s})$ for the Poisson case:

- Integration will be complicated... I cannot solve it.
- Let's consider the first two moments.
  - Mean

$$
\begin{aligned}
E(Y(\mathrm{s})) &= E(E(Y(\mathrm{s})|w(\mathrm{s}))) \\
&= E(exp(\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s}))) \\
&= e^{\mathrm{X}(\mathrm{s})'\boldsymbol{\beta}} E(e^{w(\mathrm{s})}) \\
&= e^{\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + \sigma^2/2}
\end{aligned}
$$

  - Variance

$$
\begin{aligned}
Var(Y(\mathrm{s})) &= VE(Y(\mathrm{s})|w(\mathrm{s})) + EV(Y(\mathrm{s})|w(\mathrm{s})) \\
&= V(e^{\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s})}) + E(e^{\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s})}) \\
&= V(e^{\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + w(\mathrm{s})}) + e^{\mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + \sigma^2/2}
\end{aligned}
$$

    Overdispersion

- $Cov(Y(\mathrm{s}), Y(\mathrm{u}))$ will be more complicated...

# How to fit the model

- MLE?
  - Try the likelihood
  - I cannot solve it... approximation/computing?
- Bayesian inference
  - Straightforward but MCMC is needed and parameters including $\beta$ may need to be updated with MH
  - A special case is the probit regression model for binary data

# Spatial probit model for binary data

- Non-spatial probit regression:

$$\text{Prob}(Y_i = 1) = \Phi(\mathrm{X}_i\boldsymbol{\beta})$$

  where $\Phi(\cdot)$ is the CDF of standard normal.
- Introducing the latent variable $Z_i$ such that

$$Y_i = I(Z_i > 0)$$

  where

$$Z_i \sim N(\mathrm{X}_i'\boldsymbol{\beta}, 1)$$

- Why should we set $\sigma^2 \equiv var(Z_i)$ to be 1? Identifiability! We would not be able to separate $\boldsymbol{\beta}$ and $\sigma$
- MCMC imputes $Z_i$ in each iteration. Conditioning on $Z_i$'s, it is equivalent to our classical regression with Gaussian assumption.
- Spatial version?

We set

$$Z(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + \sqrt{r}w(s) + \sqrt{(1-r)}\epsilon(\mathrm{s})$$

where

- $w(\mathrm{s})$ is GP with mean zero and variance 1 and (say) Matérn covariance function
- $\epsilon(\mathrm{s})$ iid $N(0,1)$
- Most parameters are conjugate, but we need MH for covariance parameters.
- Not everyone includes nugget in the model.

## Summary

- Non-Gaussian spatial data

## Preview:

- Nonstationary models