Spatio Temporal Data Analysis

Week 8: Modeling Areal/Lattice Data

이전에 살펴본 areal 데이터는 연속형을 가정하였다. 즉, response variable y에 대해 gaussian process을 가정해도 큰 무리가 없었다. 하지만 y가 꼭 연속형일 필요는 없고 binary, count 데이터일 수도 있다. point reference 데이터에서 y을 exponential family로 확장했었던 것처럼 areal 데이터에서도 y을 exponential family로 확장한다. 살펴볼 모델은 아래와 같다.

- Autologistic model
- Autopoisson model
- SGLMMs for areal data

Autologistic model

areal 데이터이면서 y가 binary인 경우, y의 log odds을 모델링하기 위한 방법을 살펴본다. 데이터의 예시는, 특정 동물 종이 지역별로 서식하는지의 여부를 들 수 있다. 이러한 데이터는 spatial dependence가 중요하게 작용할텐데, 만약 이를 무시하고 일반 logistic regression을 한다면 데이터를 잘 적합하지 못할 가능성이 높다.
autologistic model은 spatial dependence을 아래와 같이 모델에 통합한다.

$$P(Y_i = 1) = \frac{exp(\eta_i)}{1 + exp(\eta_i)},\ \eta_i = \beta_0 + \beta_1 X_i + \gamma Y_i^*$$

$$where\ Y_i^* = \sum_{j=1}^{n} Y_j I(i \sim j) = \sum_{j:j\sim i} Y_j$$

여기서 Y_i^*가 의미하는 바는, 지역 i의 이웃의 response 값을 모두 더한 것이다. 즉, 이를 통해 spatial dependence을 나타내고자 한다. 여기서 모수를 $\theta = (\beta_0, \beta_1, \gamma)$라고 하자. Brook's Lemma를 통해 아래의 exponential family을 얻는다.

$$P_\theta = c(\theta)^{-1} exp\left(\theta' t(y)\right)$$

$$t(y) = \left(\sum Y_i, \sum X_i Y_i, \sum Y_i Y_i^*/2\right) : sufficient\, statistics$$

- $c(\theta) = \sum_{\forall y} e^{\theta t(y)}$: intractable normalizing constant이다. 모든 y에 대해서 봐야하는데, y가 binary인 점을 고려하면, $n = 100$일 때, 고려해야할 수가 2^{100}이기 때문에 사실상 계산이 불가능하다.

- 이러한 intractable normalizing constant하에서 MLE을 어떻게 찾을까? 베이즈 접근법과 가능도 접근법이 있다.

ML Approach for Autologistic Model

- Maximum Pseudolikelihood (MPL)
  MPL은 full conditionals을 이용하여 likelihood을 근사하는 방법이다.

  $$P(Y_1, \cdots, Y_n) \approx \prod P(Y_i \mid Y_j, j \neq i)$$

  자세히 보면 근사 기호인 $\approx$을 사용하였다. Brooks Lemma는 full conditionals을 이용하여 정확한 결합 분포를 유도하는데 MPL은 근사를 한다. 즉 다시 말해서 spatial dependence을 무시하는 것이기 때문에 spatial dependence가 강하다면 bias가 많이 발생한다.

- Monte Carlo Maximum Likelihood (MCML)
  Point reference 데이터에서 살펴본 것과 유사하다. (importance sampling을 통한 적분의 근사)

  $$\begin{aligned} c(\theta) = \int exp\left(\theta' t(y)\right) dy &\propto \int \frac{exp\left(\theta' t(y)\right)}{c(\tilde{\theta})} dy \\ &= \int \frac{exp\left(\theta' t(y)\right)}{c(\tilde{\theta})} \times \frac{exp\left(\tilde{\theta}' t(y)\right)}{exp\left(\tilde{\theta}' t(y)\right)} dy \\ &= \int \frac{exp\left(\theta' t(y)\right)}{exp\left(\tilde{\theta}' t(y)\right)} \times \frac{exp\left(\tilde{\theta}' t(y)\right)}{c(\tilde{\theta})} dy \end{aligned}$$

  $$\therefore c(\theta) \approx \frac{1}{m} \sum_{i=1}^{m} \frac{exp\left(\theta' t(y_i)\right)}{exp\left(\tilde{\theta}' t(y_i)\right)},\ y_1, \cdots, y_m \sim \frac{exp\left(\tilde{\theta}' t(y_i)\right)}{c(\tilde{\theta})}$$

  $\tilde{\theta}$은 근사를 통해서 MLE와 가까워야 한다. point reference와 마찬가지로 계산이 expensive하다.

Autopoisson Model

response variable이 binary인 것뿐만 아니라 count일 수도 있다. 예를 들어 우한 바이러스의 지역별 감염자 수를 생각해보자. 후베이성의 감염자와 우리나라의 감염자수는 비슷하지 않을 것이며, 그 주변 지역과의 연관성도 다를 것이다. 바로 이러한 공간 정보를 무시하고 평범한 log linear 모델인 $Y_i \sim Poisson \sim (\mu_i),\ log(\mu_i) = X_i\beta$을 적용한다면 데이터를 잘 적합하지 못할 것이다.
Autopoisson Model은 spatial dependence을 모델에 아래와 같이 통합한다.

$$log\mu_i = X_i\beta + \sum_{j:j\sim i} \gamma_{ij} Y_j,\ with\, \gamma_{ij} < 0$$

γ_{ij} 는 이웃의 직접적인 영향력과 이웃과 공유되는 common (unmeasured) covariates을 잡아낸다. Ferrandiz에 의해서 제안된 autopoisson model의 구체적인 모습은 아래와 같다.

$$log\mu_i = X_i\beta + logu_i + \sum_{j:j\sim i} \gamma a_{ij} Y_j$$

$$where\ a_{ij} = \frac{\sqrt{u_i u_j}}{d_{ij}}, d_{ij} = distance\,between\,centroids\,of\,i, jth\,counties$$

u_i 는 인구가 많으면 count도 많아지는 것을 보정하기 위한, offset term이다. 즉, 각 지역의 인구수이다. a_{ij} 는 두 i, j 지역 간의 사람들 flow을 나타내는 인덱스이다. 만약 거리가 짧다면 a_{ij} 가 커지므로 두 지역간의 interaction effect가 커질 것이다.

- 모수는 $\theta = (\beta, \gamma)$ 이다.
- neighborhood 구조는 다른 방법으로 가정된다. 예를 들어, $i \sim j$ if $a_{ij} > \delta$. 실전에서는 δ 도 추정될 수 있지만 추론을 방해한다고 한다.
- 전과 마찬가지로 MPL과 같은 방법으로 likelihood을 근사하거나, MCML을 사용한다.

SGLMMs

Hierarchical Structure for SGLMMs for CAR / SAR Model

areal 데이터에서 SGLMMs을 하기 위한 계층 모형은 아래와 같다.

$$\mathbf{Y} \mid \mathbf{Z}, \beta \sim f\left(g^{-1}\left(\mathbf{X}\beta + \mathbf{Z}\right)\right)$$

$$\mathbf{Z} \sim N\left(0, \left(\mathbf{D}_w - \rho\mathbf{W}\right)^{-1}\right)$$

$$\rho \sim U(0, 1)$$

$\mathbf{W}$ 은 $\mathbf{Z}$ 의 neighborhood 구조를 나타내는 adjacency matrix이다. 위 계층 모형을 보면, point referenced 데이터와 공분산 구조에서 차이를 보인다. areal 데이터에 위 계층 모형을 적용하기 위해, R의 CARBayes 패키지를 사용하면 된다. 분석 과정은 크게 세 단계로 구성된다.

- 가장 중요한, neighboring structure을 만드는 단계이다. 즉, 행렬 $\mathbf{W}$ 을 만든다.
- SGLMMs을 적합한다.
- 잔차 분석을 통해, spatial model을 사용하는 것이 적절한지 확인한다.

point referenced 데이터와 마찬가지로 n 이 커짐에 따라서 computational challenges가 존재한다. 하지만 point referenced 데이터보다는 덜 심각하긴한데, 그 이유는 areal 데이터에서는 sparse precision matrix을 construct하기 때문이다.

Hierarchical Structure for SGLMMs for ICAR Model

ICAR 모델에서는 CAR / SAR과는 다르게 $\rho = 1$을 가정하였다. 따라서 $\mathbf{Q} = \mathbf{D}_w - \mathbf{W}$라고 두고 계층 모형을 아래와 같이 세운다. 앞서, $\mathbf{Q}$는 precision matrix로써, singular하기 때문에 improper한 분포를 형성한다. 만약 데이터를 모델링하는 분포가 improper하다면 문제가 되겠지만 posterior가 proper하다면 이를 prior로 써도 된다. 따라서 여기서는 improper한 분포를 prior로 쓴다.

$$\mathbf{Y} \mid \mathbf{Z}, \beta \sim f\left(g^{-1}\left(\mathbf{X}\beta + \mathbf{Z}\right)\right)$$

$$p(\mathbf{Z} \mid \tau) \propto \tau^{rank(\mathbf{Q})} exp\left(-\frac{\tau}{2}\mathbf{Z}'\mathbf{Q}\mathbf{Z}\right)$$

$$\tau^2 \sim InverseGamma(a, b)$$

전과 마찬가지로 $\mathbf{W}$는 adjacency matrix이고 이는 $\mathbf{Z}$의 neighborhood structure을 나타낸다.

Dimension Reduction Approach

point referenced 데이터에서는 공분산 행렬에 PCA를 하고, 처음 m개의 eigen vector을 취했다. areal 데이터에서는, 우선 Moran's Basis을 계산하여 $n \times n$ 행렬을 만든다. 이제 처음 m개의 principal basis을 취한다. eigen value 순서로 나열했을 때, 처음 m개의 eigen vectors을 취하여 $\mathbf{M} \in \mathbb{R}^{n \times m}$ 행렬을 만든다.
이제 $\mathbf{Z}$을 $\mathbf{M}\delta$로 대체한다. 여기서 $\mathbf{M}$은 앞서 계산한 $n \times m$ 차원의 m개의 n차원 Moran's basis로 이루어진 행렬이다. 계층 구조는 아래와 같다.

$$\mathbf{Y} \mid \delta, \beta \sim f\left(g^{-1}\left(\mathbf{X}\beta + \mathbf{M}\delta\right)\right)$$

$$p(\delta \mid \tau) \propto \tau^{rank(\mathbf{M}'\mathbf{Q}\mathbf{M})} exp\left(-\frac{\tau}{2}\delta'\mathbf{M}'\mathbf{Q}\mathbf{M}\delta\right)$$

$$\tau^2 \sim InverseGamma(a, b)$$

이제 δ의 차원은 $m << n$이므로 차원을 축소하였다.