Spatio Temporal Data Analysis

Week 5: Bayes Approach for SGLMMs

Review of Spatial Hierarchical Models When Y is Continuous

지난번에 $\mathbf{Y}$가 연속형일 때, spatial dependence가 있는 데이터에 대해서 베이지안 계층 모형을 세우는 것을 살펴보았다. 이를 간략히 리뷰해보자. $\mathbf{Y}$가 연속형일 때는, $\mathbf{Y} \mid \eta, \theta$를 GP로 가정하여, 관련된 조건부 분포나 결합 분포가 모두 GP가 되어 쉽게 conjugate 분포를 유도할 수 있었다. 이는 아래와 같다.

$$\mathbf{Y}(s_i) \mid \eta(s), \tau^2 \sim N(\eta(s_i), \tau^2) \ (Data\ Model)$$

$$\eta(s) \mid \beta, \sigma^2, \rho \sim GP(X(s)^{'}\beta, \sigma^2 K(\cdot, \cdot; \rho) \ (Process\ Model)$$

$$\beta \sim N(\mathbf{m}_\beta, \mathbf{V}_\beta)$$

$$\sigma^2 \sim Inv - Gamma(a_{\sigma^2}, b_{\sigma^2})$$

$$\tau^2 \sim Inv - Gamma(a_{\tau^2}, b_{\tau^2})$$

$$\rho \sim Gamma(a_\rho, b_\rho) \ (Parameter\ Model)$$

여기서 이전에 우리는 $K(s_i, s_j; \rho)$ 함수를 exponential covariance function으로 가정하였다.
하지만 $\mathbf{Y}$가 꼭 연속형이라는 보장은 없다. $\mathbf{Y}$가 count data일 수도 있으며, categorical data일 수도 있다. 만약 spatial denpendence를 가정하지 않는다면, GLMM을 쓰지만, spatial dependence를 가정한다면 Spatial GLMM을 사용한다. 이제부터 Spatial GLMM을 살펴보자.

Spatial Generalized Linear Mixed Models

SGLMM은 GLMM과 비교하여 Spatial이 추가되었다. 즉, GLMM이 가지는 random effect가 spatial dependence인 것이다. 그리고 $\mathbf{Y}$가 연속형, 범주형, count 데이터 중 하나임을 가정한다. 이는 GLMM에서와 마찬가지로 link function을 통해서 결정한다. 그런데 $\mathbf{Y}$가 연속형일 때와는 다르게, 한 가지 심각한 문제점이 나타난다. $\mathbf{Y}$가 연속형일 때, 세웠던 계층 모형에 대해서 다시 한번 생각해보자.

$$\mathbf{Y} \mid \eta, \tau^2 \sim Normal$$

$$\eta \mid \beta, \sigma^2, \rho \sim Normal$$

둘 모두 정규분포이므로 두 분포를 곱하여 만든 결합 분포 $\mathbf{Y}, \eta \mid \beta, \tau^2, \sigma^2, \rho$도 GP를 따르고, 여기서 η를 GP의 성질을 이용하여 integrate out함으로써 $\mathbf{Y} \mid \beta, \tau^2, \sigma^2, \rho$를 얻고, 결과적으로 사후 분포인 $\beta, \tau^2, \sigma^2, \rho \mid \mathbf{Y}$에서 표본을 얻는다. 하지만 $\mathbf{Y}$가 연속형이 아닐 때는, 이러한 관계가 성립하지 않으므로 문제가 복잡해지는 것이다. 다시 말해서, η를 integrate out해야 하는데 만약 데이터의 갯수가 1000개라면 η의 차원도 1000이 되어서 1000번의 적분을 해야하는 문제가 발생한다. 이런 문제점은 밑에서 살펴볼 Nimble을 통해서 해결한다.

SGLMM을 세울 때 발생하는 어려운 점에 대해서 살펴보았다. 이제 $\mathbf{Y}$가 count, binary 데이터일 때 어떤 식으로 모형이 세워지는지 살펴보자.

- Poisson log-linear geostatistical models

  1. $Y(s) \mid Z(s), \beta \sim Pois(exp(\eta(s)))$, $\eta(s) = X(s)^{'}\beta + Z(s)$
     이는 $Y(s)$의 conditional mean structure가 fixed effect인 β와 random effect인 $Z(s)$로 구성된다고 가정하는 것이다. 또한 log-linear라는 뜻은, 로그를 취할 때, linear하다는 뜻인데 $E[Y(s) \mid Z(s)] = exp(\eta(s))$에서 양 변에 log를 취하면 우변이 linear term으로 되기 때문이다.
  2. $Z(s) \mid \sigma^2, \rho \sim N(0, \sigma^2\Gamma(\rho))$
     random effect $Z(s)$에 대해서는 여전히 GP 가정을 한다. spatial dependence를 나타내기 위해서, $\sigma^2\Gamma(\rho)$는 exponential covariance function 등을 사용한다.
  3. 나머지는 parameter model이다.

$$\beta \sim N(0, 100\mathbf{I})$$

$$\sigma^2 \sim Inv - Gamma(0.2, 0.2)$$

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

- Binary logistic-linear geostatistical Models

  1. $Y(s) \mid Z(s), \beta \sim Ber(exp(\eta(s))/(1 + exp(\eta(s)))$, $\eta(s) = X(s)^{'}\beta + Z(s)$
     이 또한 $Y(s)$의 conditional mean structure가 fixed effect와 random effect로 구성된다고 가정하는 것이다. 또한 logistic-linear라는 뜻은 logistic 함수를 취할 때, linear하다는 뜻이다. 즉, $logit(E[Y(s) \mid Z(s), \beta]) = \eta(s) = X(s)^{'}\beta + Z(s)$이다.
  2. random effect $Z(s)$에 대한 process model도 위와 동일하게 설정한다.
  3. parameter model도 위와 동일하게 설정한다.

Nimble

$\mathbf{Y}$가 연속형이 아닐 때, MCMC가 복잡해진다고 했는데, 이를 해결해주는 R 패키지인 Nimble을 살펴보자. Nimble은 계층 모형과 같이, 계산이 많이 드는 베이지안 모형을 C++를 통해서 빠르게 계산해주는 R 패키지이다. MCMC를 직접 코딩하면서 발생할 수 있는 실수, proposal density 설정 문제 등을 자동으로 해결해준다. 초반에는 직접 코딩하면서 MCMC에 대한 이해도를 높이고, 나중에는 efficiency를 위해서 Nimble을 사용해보자.