Spatio Temporal Data Analysis

Week 4: Hierarchical Spatial Models with Bayes Approaches

Basic Setup

베이지안 통계에서는 현실의 복잡한 현상에 대해서 계층 모형 (Hierarchical Model)을 세움으로써 이를 설명하려고 한다. 대표적인 계층 모형은 Latent Dirichlet Allocation (LDA)가 있다. 간단한 모형은 사전분포를 가정하고 데이터로부터 사후 분포를 유도하면 되지만 가정해야할 사항이 많은 복잡한 모형은 계층 모형을 많이 사용한다.
앞서, 공간 데이터를 variogram을 통한 GLS를 구하거나 likelihood을 세워서 MLE를 구하는 방법을 살펴보았다. variogram을 이용하여 GLS를 구하는 방법에는, uncertainty을 control하지 못하는 단점이 있다. 즉, 추정한 variogram이 '좋은' 추정치인지 알아볼 방법이 없다. variogram을 추정할 때, 표준오차를 같이 구할 수 없기 때문이다. 하지만 베이즈 통계를 이용한다면 MCMC를 통해서 사후 분포에서 추출된 것으로 간주할 수 있는 표본을 얻으며 이를 통해 posterior variance, credibla interval을 구하여 uncertainty을 control할 수 있다.

공간 통계에서 계층 모형을 어떤식으로 specify 하는지 알아보자. 우선, 공간 통계는 관측된 데이터를 있는 그대로 바라보는 것이 아니라 이 데이터에 underlying process가 있다고 생각한다. 이 latent process는 spatial dependence와 관련된다. 관측된 데이터에 대해서 latent process가 있고, 관측되지 않은 데이터를 예측할 때, 이와 관련된 또 다른 latent process가 있다고 생각한다. 이를 notation으로 정리해보자.

- $\mathbf{Y}$는 관측된 데이터, η는 관측되지 않은 latent process, θ는 모수로 정한다.
- 계층 모형을 아래와 같이 specify한다.

$$f(\eta, \theta \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \eta, \theta) \times f(\eta \mid \theta) \times \pi(\theta) \tag{1}$$

- $f(\mathbf{Y} \mid \eta, \theta)$는 data model (likelihood), $f(\eta \mid \theta)$는 pocess model (prior), $\pi(\theta)$는 parameter model (hyperprior)이다.

η에 대해서 생각해보자. η는 관측되지 않은, latent process로, 각 관측치마다 존재한다고 가정한다. 공간 데이터를 계층 모형을 이용하여 모델링할 때, η를 모수로 간주하여 모델의 모수인 θ와의 joint posterior 분포를 고려한다; $f(\eta, \theta \mid \mathbf{Y})$ 꼴인데, 베이즈 통계에서 $f(unobserved \mid observed)$의 형태와 동일하다. 이는 다시 말하면, η는 관측된 데이터 수만큼의 차원을 가진다. 만약 $n = 1000$개의 데이터가 관측되었다면 $\eta \in \mathbb{R}^{1000}$이며 사후 분포의 차원은 $f(\eta, \theta \mid \mathbf{Y}) \in \mathbb{R}^{1000+d},\ where\ \theta \in \mathbb{R}^d$로 매우 커진다. 차원이 높아지면 MCMC가 불안정해지는 등의 문제가 발생하여 이와 관련된 연구도

활발하다고 한다. 이에 대한 해결책은 special MCMC를 이용하는 것인데, 나중에 살펴본다.
이제 η가 대충 어떤 것을 의미하는지 감이 왔을 것이다. 그런데, 관측된 데이터 뿐만 아니라 관측되지 않은 데이터에 대해서도 latent process를 가정한다고 했다. 이것이 무슨 뜻일까? 예를 들어 관측된 공간 데이터를 기반으로, 모형을 세운 후에 관측되지 않은 지역의 값을 예측하는 상황을 생각해보자. 쉽게 생각하면 신촌역과 홍대역에서 미세먼지를 관측했지만 신촌역과 홍대역 사이에있는 연희동, 연남동 등에서는 미세먼지를 관측하지 않았다면, 신촌역, 홍대역에서 관측한 미세먼지와 공간 정보를 이용하여 연희동의 미세먼지를 예측하고 싶을 것이다. 이러한 상황에서, 신촌역, 홍대역에서 관측된 미세먼지 양과 관련된 latent process, η_{obs}가 있다고 가정하고 뿐만 아니라 연희동의 예측 미세먼지 양과 관련된 latent process, η_{pred}도 있다고 가정하는 것이다.
보통 η는 $\eta(s)$로 표기하여 공간 정보 (coordinates 등)에 의존한다고 생각한다. 또한 η는 공간에서 smoothly 변화한다고 가정한다. 즉, 두 위치의 거리가 멀어질수록 spatial dependence가 줄어든다고 가정한다.

Specifying Hierarchical Model

본격적으로 모델을 specify 해보자. 우선 process model, $f(\eta \mid \theta)$을 살펴보자. process model은 아래와 같이 가정한다.

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

(2)에서 $X(s)$는 fixed mean trend이다. 이는 s에 depend하는데, 위치에 따라서 값이 달라지기 때문이다. $X(s)$는 절편과 위도, 경도 등의 공간 정보, 다른 covariates를 포함한다. (2)에서 공분산 행렬은 다음과 같다.

$$K(s_i, s_j\rho) = exp\{-||s_i - s_j||/\rho\} \tag{3}$$

(3)은 exponential covariance matrix이다. 이는, 두 공간의 거리가 멀어질수록 ($||s_i - s_j||$ 값이 커질수록) 더 작은 값을 가지는데, 이는 곧 두 공간의 correlation이 줄어듦을 의미한다. 또한 latent process에 대해

$$\eta_{obs} = (\eta(s_1), \cdots, \eta(s_n))^{'}, \ \eta_{pred} = (\eta(s_1^*), \cdots, \eta(s_m^*))^{'}$$

즉, 관측 데이터는 n개, 관측되지 않은 데이터는 m개가 있고 총 $n+m$차원의 η를 가정한다.
(2)를 보면, $\eta(s) \mid \cdot$를 GP로 가정하였다. 이는 우선 연속형 데이터만 다룰 것이기 때문이며 non GP 데이터인 count data, binomial data 등에 대해서는 후에 Spatial GLM 모델에서 살펴본다. 어찌됐든, $\eta(s) \mid \cdot$을 GP로 가정했기 때문에 관련 조건부 분포도 GP를 다른다.

$$f(\eta \mid \beta, \sigma^2, \rho) = f(\eta_{obs}, \eta_{pred} \mid \beta, \sigma^2, \rho) = f(\eta_{obs} \mid \beta, \sigma^2, \rho) f(\eta_{pred} \mid \eta_{obs}, \beta, \sigma^2, \rho) \tag{4}$$

사실 (2)에서 명시한 $\eta(s)$는 $\eta_{obs}(s)$이다. 또한 variogram을 통해 EDA를 하고, isotropic 가정을 한다면 $K(s_i, s_j; \rho) = K(h; \rho)$로 쓸 수 있다.
하지만 우리는 η_{obs}를 관측할 수 없다. η_{obs}의 첨자 obs는 η가 관측됐다는 것이 아니라, 관측된 데이터에 대한 latent process라는 뜻이다. 우리는 $Y(s_i)$를 관측하고, 이를 latent process의 noisy version이라고 간주한다. measurement error와 함께 관측되었다고 보는 것이다. $Y(s_i), \cdots, Y(s_n)$은

서로 dependent하다. 왜냐하면 spatial dependence가 존재하기 때문이다. 하지만 이러한 dependence가 주어진다면 서로 독립이라고 가정한다. 즉, $Y(s_1) \mid \eta(s_1), \cdots, Y(s_n) \mid \eta(s_n)$ 은 conditionally 독립이다. 이를 고려하여 (1)의 data model을 아래와 같이 specify 한다.

$$Y(s_i) \mid \eta(s), \tau^2 \sim N(\eta(s_i), \tau^2) \tag{5}$$

τ^2는 measurement error의 variance이다.
이제 prior를 아래와 같이 지정한다.

$$\beta \sim N(m_\beta, V_\beta)$$

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

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

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

ρ에 대해서는 conjugate prior가 없고 closed form이 나오지 않으므로 Gamma 분포를 가정한다.

Getting posterior samples through MCMC

자, 이제 prior, data model 모두 지정했으니 MCMC를 통해서 표본을 얻는 일만 남았다. 조건부 분포를 유도할 수 있다면 Gibbs, 유도할 수 없다면 MH를 쓴다고 했다. 몇 가지 유도 과정을 거치면 β, σ^2, τ^2에 대해서는 조건부 분포를 유도할 수 있다. 즉,

$$\beta \mid \sigma^2, \tau^2, \rho, \mathbf{Y} \sim Normal$$

$$\sigma^2 \mid \beta, \tau^2, \rho, \mathbf{Y} \sim Inv - Gamma$$

$$\tau^2 \mid \beta, \sigma^2, \rho, \mathbf{Y} \sim Inv - Gamma$$

하지만 ρ에 대해서는 closed form이 나오지 않는다. 따라서 MH를 통해서 표본을 얻는다.

$$\rho \mid \beta, \sigma^2, \tau^2, \mathbf{Y} \sim MH$$

B번의 iteration을 수행하였다고 하자. 그러면 아래와 같이 B 세트의 posterior samples을 얻는다.

$$\beta^{(1)}, \sigma^{2(1)}, \tau^{2(1)}, \rho^{(1)}$$

$$\cdots$$

$$\beta^{(B)}, \sigma^{2(B)}, \tau^{2(B)}, \rho^{(B)}$$

이제 B개의 표본을 가지고 베이지안 추론을 하면 된다. 예를 들어, $\frac{1}{B}\sum \beta^{(j)}$를 통해 평균을 구할 수도 있고 $B = 1000$이라면 25번째, 975번째 표본을 잘라서 95% credible interval을 구할 수도 있다.

또한, 이전에 $\widehat{\beta}_{GLS}$을 도출할 때에는 variogram을 통해 추정한 σ^2, τ^2, ρ에 대한 uncertainty을 control 할 수 없었는데, MCMC를 통해서는 표본이 주어지므로 각 표본에 대해서 credible interval을 구하는 등, uncertainty을 control할 수 있다.

Bayesian Krigging

앞서, krigging은 주어진 데이터를 기반으로 관측되지 않은 데이터를 예측하는 것이라고 배웠다. bayesian kriggig도 이와 유사한데, 베이지안 관점에서는 posterior predictive 분포를 구하여 관측되지 않은 값을 sampling 한다. posterior predictive 분포는 아래와 같다.

$$\eta_{pred} \mid \eta_{obs}, \beta, \sigma^2, \rho, \mathbf{Y} \sim N(\mathbf{m}_{pred}, \mathbf{V}_{pred}) \tag{6}$$

$$where\ \mathbf{m}_{pred} = \mathbf{X}_{pred}\beta + \gamma(\rho)^{'}\Gamma(\rho)^{-1}(\mathbf{Y} - \mathbf{X}\beta)$$

$$\mathbf{V}_{pred} = \sigma^2\left[\Gamma_{pred}(\rho) - \gamma(\rho)^{'}\Gamma(\rho)^{-1}\gamma(\rho)\right]$$

(2)에서 GP를 가정했기 때문에, (6)의 조건부 분포도 normal이 된다. η_{pred}의 표본을 얻기 위해 MCMC를 돌려야 할까? 대답은 'No'이다. MCMC를 통해 $\eta_{obs}, \beta, \sigma^2, \rho$이 표본을 모두 얻었기 때문에 이를, $\mathbf{m}_{pred}, \mathbf{V}_{pred}$에 대입하여, 조건부 정규분포의 형태를 완성하고, rnorm을 통해서 표본을 뽑으면 된다.

Difference between Likelihood based Inference

빈도론자 입장에서 likelihood을 세운다는 것은 $\beta, \sigma^2, \rho, \tau^2$가 unknown but fixed constant라고 가정하고 $Y(s_i)$에 대한 모델을 세웠다. 즉,

$$Y(s_i) = X(s_i)\beta + e(s_i),\ where\ e(s_i) \sim (0, \sigma^2\Gamma(\rho) + \tau^2\mathbf{I}_n)$$

여기서 variogram을 통해서 σ^2, ρ, τ^2의 추정치를 구하고, $\mathbf{\Sigma}(\hat{\sigma}^2, \hat{\rho}, \hat{\tau}^2)$을 구했다. 이를 이용하여, β의 GLS을 구했었다. 여기서 두 종류의 uncertaincy가 무시된다; 1. β를 추정할 때, $\theta = (\sigma^2, \rho, \tau)$에 대한 uncertainty 2. 예측할 때 β, θ에 대한 uncertainty. 베이지안 방법에서는 이러한 uncertainty를 control한다. θ을 확률 변수로 보아, 이에 대한 n개의 표본을 MCMC를 통해서 얻는다. 따라서 β의 추정치도 하나의 θ 추정치만 넣는 것이 아니라 $\hat{\beta}(\theta_1), \cdots \hat{\beta}(\theta_n)$과 같이 n개가 나와서, 추정치의 분산과 credible interval을 구할 수 있다. 모수를 fixed constant로 보느냐, 또는 확률 변수로 봐서 이에 대한 표본을 얻느냐에 따라서 관점이 달라지는 것이다.