Spatio Temporal Data Analysis

Week 3: Review of Bayesian Statistics

Introduction

저번 Week 2 포스팅에서는 시공간 데이터를 모델링할 때, 아래의 접근법을 살펴보았다.

- $(\boldsymbol{\beta}, \sigma^2, \phi)$에 대한 log likelihood를 세우고, ϕ를 고정하고 $\boldsymbol{\beta}, \sigma^2$의 MLE를 closed form으로 구하고, 이를 likelihood에 대입하여 ϕ의 MLE를 numerical하게 구한다. 이를 profile likelihood라고 한다. 물론 $(\boldsymbol{\beta}, \sigma^2, \phi)$를 동시에 optimize할 수도 있지만, 이는 computationally expensive하다. 또한, profile likelihood나 일반 likelihood을 이용하여 구한 추정치는 수학적으로 동일하기 때문에, 계산량이 적은 profile likelihood을 이용하는 것이다.
- 좀 더 naive한 방법으로, variogram을 통해 EDA를 하고, spatio dependence을 확인하여 이를 반영한 GLS를 구하는 방법이 있다.

$$\hat{\boldsymbol{\beta}}_{OLS} \rightarrow \hat{\epsilon} \rightarrow variogram \rightarrow \hat{\boldsymbol{\beta}}_{GLS}$$

  variogram을 이용하는 방법은, variogram을 추정할 때, uncertainty, 즉 variance을 얻지 못하고 model comparison을 하지 못하는 등의 단점이 있음을 살펴 보았다.

위 접근법은 장단점이 있지만, 시공간 데이터는 생각보다 더 복잡한 형태가 많다. 이런 복잡한 형태에 대해서, 종종 hierarchical models을 이용한다. 예를 들어, y를 관측하는데, y는 어떤 underlying process인 w에 의존한다고 하자; $y \mid w$. 이들의 결합 분포는 $f(y, w)$이다. 결합분포는 다루기 쉽지 않기 때문에 많은 경우에 $f(y \mid w)f(w)$와 같이 conditional $\times$ margnal로 바꾼다. 이렇게 두 분포의 곱으로 나타내는 방법이 계층 모델링이다. 시공간 데이터에서는 종종 데이터를 생성한 underlying process에 관심이 있고, 이 process는 spatial dependence를 가진다고 가정한다. 따라서 시공간 데이터를 모델링할 때, 계층 모델이 많이 사용되고, 계층 모델과 같은 복잡한 형태는 베이지안을 이용해서 해결을 한다.

본 포스팅에서는 베이지안 통계의 밑바닥부터 다루지는 않는다. 베이즈 정리, conjugate 분포, conjugate familes 유도, likelihood와 prior의 관계 등은 이미 사전 지식으로 알고 있다고 가정한다. 여기서는 MCMC에 대해서 더 집중적으로 다룬다.

Non-Conjugate Familes

사후 분포가 conjugate form이라서 손으로 계산이 가능하고 계산도 아주 쉽게 할수 있다면 얼마나 좋을까? 사실 이런 경우는 거의 없다고 보면 된다. 현실의 복잡한 문제, 계층 모형을 세우다 보면, normalizing constant까지 알 수 있는 경우는 거의 없다. 대신에, 알고리즘을 통해서, 사후 분포와 비슷한 표본을 뽑는 방법을 사용한다. 만약에 사후 분포에서 뽑았을 것으로 기대되는 표본을 얻

었다고 생각해보자. 사후 분포의 정확한 형태는 알 수 없지만 어찌됐든 true 사후 분포에서 뽑은 표본과 비슷한 표본을 가지고 있으므로 이를 이용하여 distributional quantities를 근사할 수 있다. 예를 들어, 얻은 표본을 $\theta^{(1)}, \cdots, \theta^{(B)}$ 라고 할 때, 사후 분포의 기댓값은 $\int \theta p(\theta \mid y) d\theta \approx \frac{1}{n} \sum \theta^{(i)}$ 로 근사할 수 있다. 또한 credible interval도 sample quantile을 통해서 구할 수 있다. $\theta^{(1)}, \cdots, \theta^{(B)}$ 을 오름차순으로 정렬했다고 하자. 그러면 $\theta_{0.025}, \theta_{0.975}$ 값을 통해 95% credible interval을 근사할 수 있다. 바로 MCMC를 이용하여 사후 분포와 비슷한 표본을 얻는데, 대표적인 MCMC로는 gibbs sampler, metropolis hastings 알고리즘이 있다.

Gibbs Sampler

1. 모델의 모수가 $\theta_1, \cdots, \theta_k$ 라고 가정한다.
2. random 초기값 $\theta_1^{(1)}, \cdots, \theta_k^{(1)}$ 를 정한다.
3. iteration 횟수, B 를 정하고 $i = 2, \cdots, B$ 에 대해
   (a) $\theta_1^{(i)} \mid \theta_2^{(i-1)}, \cdots, \theta_k^{(i-1)}, y$
   (b) $\theta_2^{(i)} \mid \theta_1^{(i)}, \theta_3^{(i-1)}, \cdots, \theta_k^{(i-1)}, y$
   (c) $\cdots$
   (d) $\theta_k^{(i)} \mid \theta_1^{(i)}, \cdots, \theta_{k-1}^{(i)}, y$
4. θ_1 의 표본인 $\theta_1^{(1)}, \cdots, \theta_1^{(B)}$ 에 대해서 생각해보자. θ_1 의 사후 분포, $f(\theta_1 \mid y)$ 가 무엇인지는 몰라도, 여기서 뽑혔을 거라고 기대되는 B 개의 표본이 우리 손 안에 있는 것이다. 따라서 근사 기댓값, 중앙값, credible interval 등을 계산할 수 있다. 바로 이 부분이 bayesian inference이다.

Metropolis-Hastings Sampler

Gibbs sampler는 사실 조건부 분포의 형태를 알아야 한다. 왜냐하면, 여기서 표본을 뽑아야하기 때문이다. 하지만 종종 이 조건부 분포를 도출하기 어려울 때가 있다. 이러한 일반적인 상황에서는 Metropolis-Hastings 알고리즘이 쓰인다. gibbs 처럼 초기값에서 시작하는데, MH는 proposal density 을 선택해야 한다.

1. 초기값 $\theta^{(1)}$ 과 proposal density $q(\cdot, \theta)$ 를 정한다.
2. $i = 1, \cdots, B$ 에 대해
   (a) $q(\cdot, \theta^{(i-1)})$ 로부터 θ^* 를 뽑는다. θ^* 는 바로 표본으로 채택하지 않고 아래의 acceptance rate을 계산한다.
   (b) $r = \dfrac{L(\theta^*; y)\pi(\theta^*)q(\theta^{(i-1)}, \theta^*)}{L(\theta^{(i-1)}; y)\pi(\theta^{(i-1)})q(\theta^*, \theta^{(i-1)})}$
   (c) $min\{r, 1\}$ 의 확률로 $\theta^{(i)} = \theta^*$, 즉 θ^* 을 i 번째 표본으로 채택한다. 만약 그렇지 않으면, $\theta^{(i)} = \theta^{(i-1)}$ 이라 둔다.

3. 만약 q가 symmetric하다면 q는 약분된다.

4. 보통 $q(\cdot, \theta^{(i-1)}) \sim N(\theta^{(i-1)}, \sigma_M^2)$, 정규 분포로 대부분 정한다. σ_M^2가 중요한데, 작다면 step size가 작아져서 θ^*와 $\theta^{(i-1)}$가 비슷해지고, 따라서 acceptance rate가 높아질 것이다. acceptance rate가 높아지는 것이 좋은 것은 아닌데, local minima에 빠질 수도 있기 때문이다.

Diagnosing Convergence

MCMC로부터 얻은 표본들이 실제 사후 분포에서 얻은 표본과 비슷함을 어떻게 알아낼까? 이는 곧 MCMC의 수렴에 대해서 평가하는 문제이다. MCMC는 이전 iteration의 표본에 근거하여 새로운 표본을 뽑는다. 따라서 이전 시점과의 dependence도 무시할 수 없다. 예를 들어 iid인 표본과 서로 dependence가 있는 표본을 생각해보자.

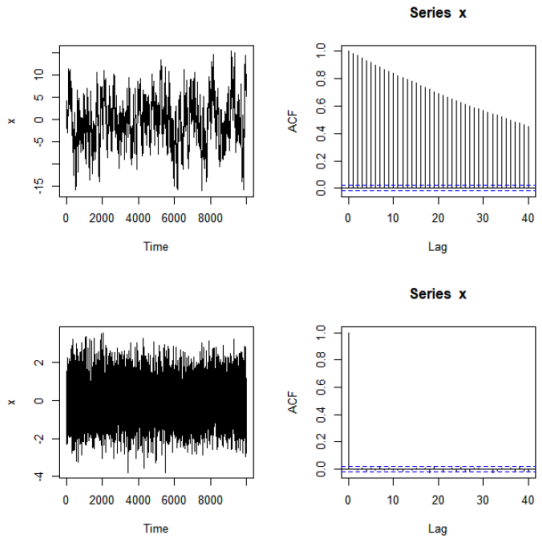
$$independence: \ \theta^{(1)}, \cdots, \theta^{(B)}$$

$$dependence: \ \psi^{(1)}, \cdots, \psi^{(B)}$$

동일한 B개의 표본이지만 dependence가 있는 $\psi^{(i)}$의 정보량은 $\theta^{(i)}$보다 더 적을 것이다. 따라서 동일한 정보량을 얻기 위해서 더 많은 표본을 뽑아야 한다. 이러한 개념이 바로 Effective Sample Size이다.

$$\widehat{ESS}(h) \propto \frac{B}{1 + 2\sum_{h=1}^{\infty} \widehat{ACF}(h)}$$

만약 lag h에서 ACF값이 있다면, 분모가 1보다 커져서 ESS가 B보다 작을 것이다.
ESS말고도 ACF의 plot을 직접 보든가, burn in period의 표본은 사용하지 않는 등의 방법이 있다. 아래 그림과 같이 graphical check를 하는 것도 한 방법이다.



아래 두 그림은 표본이 완벽한 독립인 경우이다. 현실에서는 거의 일어나지 않지만 비교를 위해 추가하였다. 아래 그림의 trace plot이 위보다 더 noisy함을 알 수 있다. 위의 trace plot은 이전 시점에

더 dependent하다. ACF에서도 큰 차이를 보인다. 아래 ACF는 독립이므로, 자기 자신과의 ACF를 제외하고는 그 값이 모두 0이다. 하지만 위에서는 ACF가 서서히 감소함을 확인할 수 있다.