Spatio Temporal Data Analysis

Week 6: ML Inference for SGLMMs

Introduction

바로 이전 포스팅에서 Bayes 모델링 시각에서 SGLMM에 접근하였다. 이는 계층 모형을 세우는 것이며 $\mathbf{Y}$가 연속형일 때와는 달리 범주형, count 데이터일 때 발생하는 문제점에 대해서도 살펴보았다. 이번에는 빈도론자 입장에서 SGLMM에 접근해본다. 즉, Likelihood를 기반으로 MLE를 구하고, 이의 근사 분포를 생각해본다.

베이지안이든 빈도론자이든, SGLMM에서는 η의 차원이 높아져서 η를 integrate out하는데 어려움이 있음을 살펴보았다. 베이지안은 이를 Nimble 패키지를 통해서 해결한다 (?) 빈도론자는 이를 어떻게 해결할까? 바로 이와 같이 η에 대한 적분 문제를 어떻게 해결하는지에 따라서 베이지안, 빈도론자의 방법의 차이점이 드러난다. ML을 이용한 SGLMM 접근법에는 크게 세 가지가 있다.

1. Monte Carlo maximum Likelihood
2. Monte Carlo expectation-maximization
3. Laplce approximation

각 방법에 대해서 자세하게 살펴보기 이전에, ML 접근법을 formal하게 정의해보자. 우리는 Likelihood $f(Y \mid \beta, \sigma^2, \phi)$를 구하고 싶고 이는 결합 분포인 $f(Y, Z \mid \beta, \sigma^2, \phi)$에서 Z를 integrate out하여 얻는다.

$$L(\psi) = \int_{R^n} f_{Y|Z}(Y \mid Z, \beta) f_Z(Z \mid \sigma^2, \phi) dZ, \ where \ \psi = (\beta, \sigma^2, \rho, \tau^2)$$

이를 최대로 만드는 것이 바로 MLE이다.

$$\widehat{\psi}_{MLE} = argmax_{\psi} L(\psi)$$

Monte Carlo Maximum Likelihood (MCML) [Christensen, 2004]

random effect와 데이터의 결합 분포를 생각해보자.

$$f_{Y,Z}(Y, Z \mid \psi) = f_{Y|Z}(Y \mid Z, \beta) f_Z(Z \mid \sigma^2, \phi)$$

또 데이터를 조건으로 하는 random effect의 조건부 분포를 생각해보자.

$$f_{Z|Y}(Z \mid Y, \psi) \propto f_{Y,Z}(Y, Z \mid \psi) = f_{Y|Z}(Y \mid Z, \beta) f_Z(Z \mid \sigma^2, \phi)$$

이를 이용하면 Likelihood는

$$\begin{aligned} L(\psi \mid Z) &= \int_{R^n} f_{Y,Z}(Y, Z \mid \psi) dZ \\ &= \int_{R^n} f_{Y,Z}(Y, Z \mid \psi) \times \frac{f_{Y,Z}(Y, Z \mid \tilde{\psi})}{f_{Y,Z}(Y, Z \mid \tilde{\psi})} dZ \\ &\propto \int_{R^n} \frac{f_{Y,Z}(Y, Z \mid \psi)}{f_{Y,Z}(Y, Z \mid \tilde{\psi})} \times f_{Z|Y}(Z \mid Y, \tilde{\psi}) dZ \end{aligned}$$

직접적으로 integrate을 하는 것이 아니라, importance sampling을 통해서 적분을 근사하는 것이 핵심 아이디어이다. $f_{Z|Y}(Z \mid Y, \tilde{\psi})$는 importance function이고 $\tilde{\psi}$는 MLE와 가깝게 세팅되어야 한다. Likelihood를 아래와 같이 두고

$$log\, L(\psi \mid Z) = l(\psi) = log \left(\int_{R^n} \frac{f_{Y,Z}(Y, Z \mid \psi)}{f_{Y,Z}(Y, Z \mid \tilde{\psi})} \times f_{Z|Y}(Z \mid Y, \tilde{\psi}) dZ \right)$$

$Z^{(1)}, \cdots, Z^{(K)} \sim f_{Z|Y}(Z \mid Y, \tilde{\psi})$을 이용해서 Likelihood에 대해서 monte carlo approximation을 한다.

$$\widehat{l}(\psi) = log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{f_{Y,Z}(Y, Z^{(k)} \mid \psi)}{f_{Y,Z}(Y, Z^{(k)} \mid \tilde{\psi})} \right)$$

MLE는

$$\widehat{\psi} = argmax_{\psi} \widehat{l}(\psi)$$

실제로는, $\tilde{\psi}$를 MLE와 가깝게 설정하는 것이 불가능하다. 왜냐하면 애초에 MLE를 모르기 때문에 이 알고리즘을 쓰는 것이기 때문이다. 따라서 iteratvie search을 통해서 $\tilde{\psi}$를 최대한 MLE와 가깝게 찾는다. 즉, ψ^0를 initial value라고 하자.

$$l(\psi) \approx log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{f_{Y,Z}(Y, Z \mid \psi)}{f_{Y,Z}(Y, Z \mid \psi^0)} \right)$$

을 최대로 하는 ψ를 ψ^1이라고 하자. iterative하게 이 과정을 반복하고, ψ^i가 $\tilde{\psi}$로 수렴하면 그 값이 MLE와 가깝다고 생각한다.
MCML은 아래의 장/단점이 있다.

- 장점
  - MCML에 대한 theoritical justification이 이미 연구되었다. 따라서 likelihood 기반 추론이 가능하다. 즉, likelihood에 대한 근사치를 구하기 때문에 AIC, BIC 등을 계산할 수 있다.
  - spatial model에 국한되지 않고, latent variable models에 다양하게 적용할 수 있다.
- 단점
  - $\tilde{\psi}$를 찾아야 한다. K개의 Monte Carlo 표본을 뽑고 n번의 iteration을 수행한다면, complexity는 $O(nk)$이다.

Monte Carlo Expectation Maximization (MCEM) [Zhang, 2002]
MCEM은 MCML과 유사하게 Monte Carlo 표본을 사용한다. 하지만 MCML은 Likelihood의 근사치를 구하고 MCEM은 EM을 통해 ψ를 근사한다는 것이 다른 점이다. $\psi^{(t)}$를 t번째 iteration의 parameter value라고 하자. SGLMM에서 관측되지 않은 것은 Z, latent process이다. EM에서는 관측되지 않은 것을 모두 condition으로 집어넣고, complete log likelihood의 conditional expectation을 구한다.

$$Q(\psi, \psi^{(t)}) = E\left[log\, f_{Y,Z}(Y, Z \mid \psi) \mid Z, \psi^{(t)}\right]$$

이후 E, M step은 usal EM 과정와 동일한데, Q를 Monte Carlo 표본을 이용하여 근사시킨다는 것만 다르다.

- E step: MCMC 표본을 $f_{Z|Y}(Z \mid Y, \psi^{(t)})$에서 뽑는다; $Z^{(1)}, \cdots, Z^{(K)} \sim f_{Z|Y}(Z \mid Y, \psi^{(t)})$ MCMC 표본을 이용하여 Q를 근사한다.
- M step: $Q(\psi^{(t+1)}, \psi^{(t)}) > Q(\psi^{(t)}, \psi^{(t)})$을 만족하는 $\psi^{(t+1)}$을 iterative하게 찾아나간다. closed form이 없으므로 first, second moment을 이용한 N-R 최적화를이용한다.

장점은 missing data analysis의 EM을 가져온 것이므로, theoritical justification을 필요로 하지 않는다. 하지만 likelihood의 근사치를 구하지 않아서 이와 관련된 추론을 할 수 없다.

Laplace Approximation [Bonat and Ribeiro, 2016]
$L(\psi \mid Y)$을 Gaussian distribution을 통해서 근사시킨다.

$$L(\psi \mid Y) = \int_{R^n} exp(T(Z))dZ \approx (2\pi)^{-n/2} \mid -\frac{\partial^2 T(\hat{Z})}{\partial^2 Z} \mid exp(T(\hat{Z}))$$

$$where\ T(Z) = log\, f_{Y,Z}(Y, Z \mid \psi)$$

$\hat{Z}$는 주어진 ψ에 대한 optimized value이고 $\hat{Z}$를 정했으면 이제 ψ에 대한 optimization을 푼다. Laplace Approximation은 계산이 가장 빠르지만 theoritical justification이 없고, 심지어 likelihood을 최대로 만들지도 않는다.

Computational Challenges for SGLMM

SGLMM에 대한 bayes approach는 latent process의 차원이 종종 매우 커서, 고차원 분포에서 MCMC 표본을 뽑는 것에 문제가 있었다. SGLMM을 위한 hierarchical structure는 아래와 같다.

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

$$Z(s) \mid \sigma^2, \rho \sim N(0, \sigma^2 \Gamma(\rho))$$

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

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

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

즉, Z의 차원이 높아서 고차원에서 표본을 뽑는 것에 어려움이 있다. 게다가 얘네들은 spatial dependence를 나타내는 것이기 때문에 종종 higly correlated 되어 있다. 서로 높은 dependence가 있는 표본들은 만개를 뽑았다고 해도, ESS를 계산해보면 훨씬 작게 나올 것이다. 따라서 매우 많은 iteration이 필요하다.
likelihood based approach에서는 $L(\psi) = \int_{R^n} f_{Y|Z}(Y \mid Z, \beta) f_Z(Z \mid \sigma^2, \phi) dZ,\ where\ \psi = (\beta, \sigma^2, \rho, \tau^2)$의 likelihood function을 구해야 하는데, Z에 대해서 적분을 해야한다. 즉, 고차원 적분을 해야하므로 계산이 오래 걸릴 것이다.

Projecting Random Effects
고차원 행렬을 저차원 행렬로 근사하는 방법은 SVD가 대표적이다. SVD의 아이디어를 빌리고, 여기에 probability 개념을 추가한 방법을 Guan, Haran이 제안하였다. notation을 정리하자.

- Z: original latent process
- U_ρ: first m eigenvectors of correlation matrix Γ_ρ $(n \times m)$
- D_ρ: a diagonal matrix with corresponding eigenvalues $(m \times m)$
- $M_\rho = U_\rho D_\rho^{1/2}$: projection matrix $(n \times m)$

이를 이용하여 저차원의 모델을 가정한다. (Y가 continous인 경우)

$$g\{E[Y \mid \beta, M_\rho, \delta]\} = X\beta + M_\rho\delta,\ \delta \sim N(0, \sigma^2\mathbf{I}),\ \mathbf{I}: 3 \times 3\ matrix$$

원래, $g\{E[Y \mid \beta, Z, \delta]\} = X\beta + Z,\ Z \sim N(0, \sigma^2\Gamma(\rho))$의 형태였던 것을 떠올려보자. Z의 matern covariance $\sigma^2\Gamma(\rho)$는 $n \times n$이다. 그런데, n이 클 경우에 여러 문제점이 발생되므로 $n >> m$인 m을 골라서 M_ρ을 통해 Z의 차원을 줄이는 것이다. 이때 사용되는 criteria는 임의로 정하는데, 보통 처음 m개의 eigenvalues을 사용한다고 하면 $\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^n \lambda_i} > 0.95$을 기준으로 m을 정한다. reduced model의 matern covariance을 살펴보자.

$$\delta \sim N(0, \sigma\mathbf{I})$$

$$\therefore M_\rho\delta \sim N(0, \sigma^2 UDU^T),\ Z \sim N(0, \sigma^2\Gamma(\rho))$$

저차원으로 projection을 통해서 아래의 계층 모형을 생각한다.

$$Y \mid \delta, \beta \sim f(g^{-1}(X\beta + M_\rho\delta))$$

$$\delta \mid \sigma^2 \sim N(0, \sigma^2\mathbf{I})$$

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

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

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

$m << n$이므로 빠른 계산을 기대할 수 있고 δ가 서로 독립이므로 적절한 크기의 ESS를 얻기 위해 많은 iteration을 할 필요도 없을 것이다.