The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.353
28(2), 353–360

Review of Spatial Linear Mixed Models for Non-Gaussian Outcomes

Jincheol Park[a,1]

[a]Department of Statistics, Keimyung University

(Received April 1, 2015; Revised April 6, 2015; Accepted April 6, 2015)

Abstract

Various statistical models have been proposed over the last decade for spatially correlated Gaussian outcomes. The spatial linear mixed model (SLMM), which incorporates a spatial effect as a random component to the linear model, is the one of the most widely used approaches in various application contexts. Employing link functions, SLMM can be naturally extended to spatial generalized linear mixed model for non-Gaussian outcomes (SGLMM). We review popular SGLMMs on non-Gaussian spatial outcomes and demonstrate their applications with available public data.

Keywords: Non-Gaussian data, spatial generalized linear mixed model.

1. 개요

관측값간의 공간적 의존관계를 고려한 회귀분석은 다양한 분야에서 응용되고 있는데, n 사이트 s_i, $i = 1, \ldots, n$에서 관측된 연속형 관측값 $Y(s_i)$에 대해서 랜덤효과(random effect) $\{Z(s_i)\}$을 모형에 포함하여 다음과 같이 선형 혼합모형으로 모형화하는 것이 일반적이다:

$$Y(s_i) = W_i^T \boldsymbol{\beta} + Z(s_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N\left(0, \tau^2\right), \tag{1.1}$$

여기서 W_i는 $Y(s_i)$의 공변량(Covariate)이며 $\boldsymbol{\beta}$는 공변량에 대응되는 회귀계수이다. $\{Z(s_i)\}$를 모형화하는 데에 있어서 Gaussian Field(GF) 모형을 가정하는 것이 일반적인데 이럴 경우에 $E\{Z(s_i)\} = 0$와 $\mathrm{Var}\{Z(s_i)\} = \sigma^2$로 주어진다. 또한 $Z(s_i)$와 $Z(s_j)$의 상관관계는 $\mathrm{Corr}\{Z(s_i), Z(s_j)\} = \rho(\|s_i - s_j\|; \boldsymbol{\theta})$로 모형화되는데 여기서 $\|\cdot\|$는 유클리드 거리이며 상관관계함수 $\rho(\,\cdot\,;\boldsymbol{\theta})$는 Matérn 또는 Spherical 모형 (Cressie, 1993)과 같이 모수적 모형 중에서 선택하게 되는데 $\boldsymbol{\theta}$는 ρ 함수와 연관되는 모수이다. 모형 (1.1)에 대하여, 랜덤벡터 $\boldsymbol{Y} = \{Y(s_1), \ldots, Y(s_n)\}^T$는 다음과 같은 다변량 정규분포를 따르게 된다:

$$\boldsymbol{Y} \sim \mathrm{MVN}\left\{\boldsymbol{W}\boldsymbol{\beta}, V\left(\boldsymbol{\theta}, \tau^2\right)\right\},$$

여기서 $\boldsymbol{W} = (W_1^T, \ldots, W_n^T)$이고 $n \times n$ 항등행렬 I에 대하여 $V(\boldsymbol{\theta}, \tau^2) = \Sigma(\boldsymbol{\theta}) + \tau^2 I$이다.

[1]Department of Statistics, Keimyung University, 2800 Dalgubeol-daero, Dalseo-gu, Daegu, Korea.
E-mail: park.jincheol@gw.kmu.ac.kr

Diggle 등 (1998)은 이 모형을 일반화된 공간선형 혼합모형으로 확장하였는데 모형 (1.1)의 $Y(s_i)$를 $E\{Y(s_i)\}$로 대체하고 연결함수(link function)를 활용하여 $\eta[E\{Y(s_i)\}] = W_i^T\boldsymbol{\beta} + Z(s_i)$로 정의한다. 이렇게 하면 두 단계 모형 또는 계층적 모형(hierarchical model)이 되는데 전체우도(full likelihood)는 다음과 같이 주어지며 모수 추정은 주로 MCMC 기법을 활용하게 된다:

$$\prod_i p\left\{Y(s_i)|z(s_i),\boldsymbol{\beta},\boldsymbol{\theta},\tau^2\right\} p\left(\boldsymbol{z}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\beta},\boldsymbol{\theta},\tau^2\right). \tag{1.2}$$

본 논문에서는 $Y(s_i)$가 이산형인 경우에 대하여 널리 활용되는 모형을 살펴보고 적용사례를 알아 보고자 한다.

2. 모형

$Y(s_i)$간의 공간적 위치에 따르는 의존성을 모형화하는데 있어서 크게 두 가지 유형으로 나눌 수 있는데 정규적(regular) 격자 위에 s_i가 위치하는 격자모형과 관심있는 지역에서 불규칙적인(irrelgular) 위치에서의 관측을 가정하는 모형이 있다.

반응변수 $Y(s_i)$가 Binary인 경우 격자 모형으로 가장 유명한 모형 중 하나로는 Besag (1974)이 제안한 Autologistic 모형이 있다. 이 모형의 특징은 $Y(s_1),\ldots,Y(s_n)$간의 상관관계를 직접적(non-hierarchical)으로 그리고 조건부적으로 정의한다는 것이다. Autologistic 모형을 살펴보면

$$\log\frac{P\{Y(s_i)=1\}}{P\{Y(s_i)=0\}} = W_i^T\boldsymbol{\beta} + \sum_{j\in N(s_i)}\eta_{ij}Y(s_j) \tag{2.1}$$

으로 주어지는데 여기서 W_i는 $Y(s_i)$의 공변량 벡터이며 $\boldsymbol{\beta}$는 대응되는 회귀계수를 나타내며 $\eta = \{\eta_{ij}\}$는 의존계수로서 Z_i의 근방(neighbor)에서만 0이 아닌 값을 취한다. $\eta_{ij} = \eta I_{i\sim j}$를 가정하고 ($\sim$은 근방관계를 나타낸다) Pairwise 의존성을 가정하면 결합분포는

$$p(Z|\boldsymbol{\theta}) = c(\boldsymbol{\theta})^{-1}\exp\left\{Z^TX\boldsymbol{\beta} + \frac{\eta}{2}Z^TAZ\right\}$$

의 형태를 띄게 된다. 여기서 A는 $n\times n$ 근방(adjacency) 행렬로써 $A_{ij} = I_{i\sim j}$이고 $c(\boldsymbol{\theta})$는 계산이 매우 어려운(intractable constant) 상수가 된다. $Q(Z|\boldsymbol{\theta})$를 $Q(Z|\boldsymbol{\theta}) = Z^TX\beta + (\eta/2)Z^TAZ$로 두면

$$p(Z|\boldsymbol{\theta}) = \frac{\exp\{Q(Z|\boldsymbol{\theta})\}}{\sum_{Y\in\Omega}\exp\{Q(Z|\boldsymbol{\theta})\}}$$

을 얻을 수 있다. Caragea와 Kaiser (2009)는 고전적 Autologistic 모형의 모수가 서로 Confound 되어 모형 식별에 문제가 있음을 지적하고 Centered Autologistic 모형을 다음과 같이 제시하였다:

$$\log\frac{P(Y(s_i)=1)}{P(Y(s_i)=0)} = W_i^T\boldsymbol{\beta} + \sum_{j\in N(s_i)}\eta_{ij}(Y(s_j)-\mu_j)$$

이며 μ_j는 다음과 같이 정의된다:

$$\mu_j = E\{Y(s_j)|\boldsymbol{\eta}=0\} = \frac{\exp(W_j\boldsymbol{\beta})}{1+\exp(W_j\boldsymbol{\beta})}.$$

Non-Gaussian 데이터 모형에 대한 더 일반적인 접근방법은 계층적구조(hierarchical structures) 방법론인데 다음과 같이 모형화 한다:

$$\eta[E\{Y(s_i)\}] = W_i^T\boldsymbol{\beta} + Z(s_i).$$

계층적 모형의 경우 대부분 MCMC 기법을 사용하여 모수를 추정하게 되는데 모형 (1.2)을 그대로 사용한다면 데이터 사이즈 n이 조금만 커져도 추정이 현실적으로 어렵게 된다. 따라서 빅데이터를 다루기 위해서 여러가지 방법론이 제시되었는데 그 중에서 일반화 모형에 사용될 수 있는 방법으로 저차원 근사법이 있다. 이 방법론은 Gaussian 공간 프로세스(Gaussian process) 즉 $Y(s_i)$가 Gaussian인 경우에 적용되는 것이 일반적이다. 스무딩(smoothing), 커널 컨벌루션(kernel convolution), 무빙 평균(moving average), 저차원 스플라인(low-rank spline) 또는 기저 근사법등을 통하여 $\{Z(s)\}$을 저차원의 $\{\tilde{Z}(s)\}$으로 근사하는 기술이다 (Wikle와 Cressie, 1999; Lin 등, 2000; Kammann와 Wand, 2003; Paciorek, 2007; Banerjee 등, 2008). 그 중에서 Banerjee 등 (2008)가 제안한 방법론을 살펴보자. 우선 매듭(knots)을 구성하는 데에 있어 관측된 사이트의 부분집합으로 구성하거나 관심 지역을 커버 할 수 있도록 구성할 수도 있다. 구성된 매듭을 $\boldsymbol{s}^* = \{s_1^*, \ldots, s_m^*\}$라 두면 Gaussian Field $\{Z\}$로부터 $\tilde{Z}^* = \{Z(s_1^*), \ldots, Z(s_m^*)\}^T$을 다음과 같이 구성할 수 있다:

$$\tilde{\boldsymbol{Z}}^* \sim \text{MVN}\left(\mathbf{0}, \tilde{\Sigma}(\boldsymbol{\theta})\right),$$

여기서 $\tilde{\Sigma}(\boldsymbol{\theta})$는 $m \times m$ 행렬이므로 $n \times n$ 행렬에 비해서 계산적인 이점이 있다. 사이트 s_0에서의 $\tilde{Z}(s_0)$는 $E[Z(s_0)|\tilde{\boldsymbol{Z}}^*]$로 정의되어 $c_0^T(\boldsymbol{\theta})\tilde{\Sigma}(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{Z}}^*$로 계산되는데 여기서 $c_0(\boldsymbol{\theta})$는

$$c_0(\boldsymbol{\theta}) = \{\text{Cov}(Z(s_0), Z(s_1^*)), \ldots, \text{Cov}(Z(s_0), Z(s_m^*))\}^T$$

이다. 부모 프로세스 $Z(\boldsymbol{s})$로 부터 비롯되는 $\tilde{Z}(\boldsymbol{s})$를 Benerjee 등 (2008)는 예측 프로세스(predictive process)라고 불렀으며 공간모형 (1.1)을 다음의 예측 프로세스로 근사하는 것을 제안하였다:

$$Y(\boldsymbol{s}) = W_i^T\boldsymbol{\beta} + \tilde{Z}^*(\boldsymbol{s}) + \varepsilon_i. \tag{2.2}$$

결과적으로 모형 (2.2)를 적합하는데에 있어서 n개의 랜덤 효과 Z를 다루기 보다는 m개의 랜덤효과 $\tilde{Z}$만 다루기 때문에 추정과정에서 $m \times m$ 공분산 행렬의 역행렬을 이용하게 되어 계산비용을 줄일 수 있다. 계층적 구조를 적용하면 Gaussian Field를 잠재층(latent layer)으로 두고 링크를 적용하여 $g[E\{Y(s)\}] = W_i^T\boldsymbol{\beta} + \tilde{Z}(s)$로 정의하면 두단계 모형 또는 계층적 모형(hierarchical model)이 되는데 전체우도(full likelihood)는 다음과 같이 주어진다:

$$\prod_i P\{y(s_i)|\tilde{\boldsymbol{z}}^*, \boldsymbol{\beta}, \boldsymbol{\theta}\}P(\tilde{\boldsymbol{z}}^*|\boldsymbol{\beta}, \boldsymbol{\theta})P(\boldsymbol{\beta}, \boldsymbol{\theta}). \tag{2.3}$$

빅데이터를 다루는데 있어 또 다른 접근방법으로 공간 프로세스를 Markov 랜덤필드(Markov Random Field)로 근사하는 방법이 있는데 널리 활용되는 기술이다 (*e.g.*, Rue와 Tjelmeland, 2002; Rue와 Held, 2005). 이 방법론은 처음에 규칙적(regular) 격자 모형에 적용되었다가 후에 불규칙적(irregularly)으로 분포하는 $Y(s_i)$로 공간모형으로 확장되었다. 예를 들면 Hartman과 Hössjer (2008)은 불규칙 공간 프로세스를 격자 위의 MRF로 근사한 후 격자 위에 위치하지 않는 사이트의 프로세스를 interpolation하는 방법을 제시하였다.

Non-gaussian에 대하여 가장 일반적으로 사용되는 모형 중 하나는 ICAR(intrinsic conditional auto regression)로서 모형 (2.2)에서 $Z(s_i)$를 평균이 0인 GMRF로 정의하는 것이다. Rue 등 (2009)는 $Z(s_i)$가 GF이거나 $Y(s)$가 Non-Gaussian인 경우에 신속한 계산을 위하여 INLA(Integrated nested Laplace approximation) 기술을 소개하였다. INLA에서 사후 확률은

$$p(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta}) \prod_{i \in N(s_i)} p(y_i|z_i, \boldsymbol{\theta})$$

로 계산되는데 $p(y_i|z_i,\boldsymbol{\theta})$가 Gaussian이 아니라면 이 확률은 계산 비용이 높다. INLA은 첫 단계로 $\boldsymbol{\theta}$의 주변 사후 확률을 다음과 같이 주어지는 $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$로 근사한다:

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{y})}{p_G(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{z}=\boldsymbol{z}^*(\boldsymbol{\theta})},$$

여기서 $p_G(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$는 주어진 $\boldsymbol{\theta}$에 대해서 $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$를 근사하고자 하는 원분포와 모드 $\boldsymbol{z}^*(\boldsymbol{\theta})$에서 매칭시켜서 구한 $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$의 Gaussian 근사이다. 이 Gaussian 근사 $p_G(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$가 가능한 이유는 $p(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$가 다음과 같은 이차식으로 표현이 가능하기 때문인다:

$$\begin{aligned} p(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{y}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{z}|\boldsymbol{\theta})\prod_{i=1}^{n} p(y_i|z_i,\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta})Q(\boldsymbol{\theta})^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\boldsymbol{z}^T Q\boldsymbol{z} + \sum_{i=1}^{n}\log p(y_i|z_i,\boldsymbol{\theta})\right\}. \end{aligned}$$

위의 식에서 $p(\boldsymbol{\theta}|\boldsymbol{y})$이 Gaussian이 아닌 경우가 일반적이므로, 사후 주변 확률의 근사는 $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$을 이용하여 수치해석적으로 계산한다.

다음 단계는 잠재계층의 사후 주변 확률 $p(z_i|\boldsymbol{y})$을 다음과 같이 구하는데,

$$\tilde{p}(z_i|\boldsymbol{y}) = \sum_k \tilde{p}(z_i|\boldsymbol{\theta}_k,\boldsymbol{y})\tilde{p}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k,$$

여기서 $\tilde{p}(z_i|\boldsymbol{\theta}_k,\boldsymbol{y})$는 $p(z_i|\boldsymbol{\theta}_k,\boldsymbol{y})$의 다음에 주어지는 식에 기반한 라플라스 근사치이다:

$$\tilde{p}(z_i|\boldsymbol{\theta},\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{y})}{p_G(\boldsymbol{z}_{-i}|z_i,\boldsymbol{\theta},\boldsymbol{y})}\right|_{\boldsymbol{z}_i=\boldsymbol{z}_i^*(z_i,\boldsymbol{\theta})},$$

여기서 $p_G(\boldsymbol{z}_{-i}|z_i,\boldsymbol{\theta},\boldsymbol{y})$는 $p(\boldsymbol{z}_{-i}|z_i,\boldsymbol{\theta},\boldsymbol{y})$에 대한 Gaussian 근사이고 $\boldsymbol{z}_i^*(z_i,\boldsymbol{\theta})$는 모드값이다. Lindgren 등 (2011)은 SPDE(stochastic partial differential equation)을 풀어서 Matérn의 공간의존성 구조를 가지는 GMRF는 GF로 표현가능함을 밝혔다. 관측값이 존재하는 불규칙적(irregular)사이트를 삼각형의 형태로 잘게 나누어서 GMRF로 근사할 수 있으므로 실제 데이터 분석에서는 GMRF의 각 격자의 근방(neighbor)을 정의하는 데에 효과적이다.

한편 Reich 등 (2006)는 모형 (1.2)에서 랜덤 효과 $\{Z(s_i)\}$가 공변량 $\{W_i\}$의 행렬공간 $\mathcal{C}(\boldsymbol{W})$와 간섭현상이 발생하게 되어 $\boldsymbol{\beta}$의 사후 분포에 편이와 분산을 증대시키는 효과가 발생한다는 사실을 발견하였다. 이를 극복하기 위하여 $\mathcal{C}(\boldsymbol{W})$로의 직교사영(orthogonal projection) P를 고려하면 $I-P$는 $\mathcal{C}(\boldsymbol{W})^{\perp}$로의 직교사영이 된다. P와 $I-P$의 직교기저 행렬 K와 L를 구하게 되면, $g[E\{Y(s_i)\}] = W_i^T\boldsymbol{\beta} + Z(s_i)$를 다음과 같이 분해할 수 있다:

$$g[E\{Y(s_i)\}] = W_i^T\boldsymbol{\beta} + k_i^T\gamma + l_i^T\delta$$

여기서 k_i^T, l_i^T는 P와 $I-P$의 기저로서 K와 L의 행벡터이며 γ와 δ는 랜덤계수가 된다. 이 식에서 알 수 있는 것은 K와 W가 동일한 열 공간을 가지게 되어 모형의 유일성(identifiability)가 보장되지 않는다는 것이다. 따라서 Reich 등 (2006)는 $k_i^T\gamma$를 제거한

$$g^*[E\{Y(s_i)\}] = W_i^T\boldsymbol{\beta} + l_i^T\delta$$

으로 모형화할 것을 제안하였고 Hughes와 Haran (2013)은 더욱 발전시켜 sparse SGLMM(Spatial GLMM)을 제안하였다.

Table 3.1. Explanation about variables

변수	변수 설명
DEATH	영아사망자 수
BIRTHS	생존 출생자 수
LOW	저체중아 비율
BLACK	흑인 거주 비율 (2000 US Census)
HISP	히스패닉 거주 비율 (2000 US Census)
GINI	지니 계수
AFF	사회적 부유계수
STAB	거주 안정성 (2000 US Census)

3. 데이터 분석

이번 절에서는 R에 구현되어 있는 패키지들을 활용하여 실제로 데이터 분석을 어떻게 진행할 수 있는지를 설명하고자 한다. 분석할 데이터는 2008년 ARF(Area Resource File)인데 이는 미국의 카운티 수준의 데이터베이스로서 3071개의 카운티에서 조사한 영아사망율(infant mortality)이다. 2002년부터 2004년까지 3년간 영아사망 변수로 구성되어 있으며 그 외의 변수는 US census 등 다른 데이터베이스에서 가져온 변수이다. 이 데이터는 ngspatial이라는 R-package (Hughes와 Cui, 2015)에서 가용하다. 변수에 대한 정보는 Table 3.1에 설명되어 있다.

다음과 같은 Poisson 모형으로 적합하였는데 $Z(s_i)$은 ngspatial 패키지에 구현되어 있는 sparse SGLMM과 INLA R-패키지 (Rue 등, 2014)에 구현되어 있는 ICAR 그리고 $Z(s_i) \overset{iid}{\sim} N(0, \sigma_z^2)$를 가정하는 독립모형(IID)으로 모형화하여 비교하였다:

$$\log\{E(Y(s_i)|\boldsymbol{\beta}, \boldsymbol{\theta})\} = \log(\text{BIRTH}_i) + \beta_0 + \beta_1 \text{LOW}_i + \beta_2 \text{BLACK}_i + \beta_3 \text{HISP}_i + \beta_4 \text{GINI}_i + \beta_5 \text{AFF}_i + \beta_6 \text{STAB}_i + Z(s_i).$$

분석에 사용된 코드는 다음과 같으며 그 적합 결과는 Table 3.2에 종합하였다.

```
library(ngspatial)
data(infant)
infant$low_weight = infant$low_weight / infant$births
attach(infant)
Z = deaths
X = cbind(1, low_weight, black, hispanic, gini, affluence, stability)

###### Sparse SGLMM
m1= sparse.sglmm(Z ~ X - 1 + offset(log(births)), family = poisson,
A = A,tune = list(sigma.s = 0.02), verbose = TRUE)

ID=1:nrow(X)
INF=data.frame(Z,X)
###### INLA with ICAR
m2<-inla(Z ~ X - 1 + offset(log(births))+f(ID, model="besag",
graph=A,adjust.for.con.comp = FALSE),
```

Table 3.2. Results of parameters estimation. Numbers in parenthesis mean two limits of 95 % credible interval.

변수	sparse SGLMM	ICAR	IID
LOW	8.8040(7.567, 10.050)	7.7676(6.401, 9.128)	8.3275(7.051, 9.598)
BLACK	0.0042(0.002, 0.005)	0.0040(0.002, 0.005)	0.0043(0.003, 0.005)
HISP	−0.0039(−0.004, −0.002)	−0.0032(−0.004, −0.001)	−0.0038(−0.004, −0.002)
GINI	−0.5526(−0.980, −0.123)	−0.0796(−0.555, 0.399)	−0.4756(−0.931, −0.018)
AFF	−0.0756(−0.087, −0.063)	−0.0773(−0.091, −0.063)	−0.0824(−0.095, −0.069)
STAB	−0.0283(−0.043, −0.013)	−0.0420(−0.059, −0.024)	−0.0355(−0.051, −0.019)

```
family="poisson", data=INF,
control.predictor=list(compute=TRUE))

###### INLA with IID
m2<-inla(Z ~ X - 1 + offset(log(births))+f(ID, model="iid",
graph=A,adjust.for.con.comp = FALSE),
family="poisson", data=INF,
control.predictor=list(compute=TRUE))
```

각각의 모형에 의해 추정된 공간 효과 $Z(s_i)$의 $\hat{\sigma}_Z$는 각각 0.33, 0.097, 0.068이였다. Table 3.2에 의하면 독립모형(IID)과 sparse SGLMM에서는 모든 변수가 유효한 것으로 나타나는 반면 ICAR모형에서는 GINI 변수를 제외한 모든 변수에서 유효한 것으로 나타난다. Reich 등 (2006)에서 지적되었듯이 sparse SGLMM과 ICAR 모형 간에 주목할 만한 차이는 HPD 구간에서 알 수 있듯이 ICAR의 회귀계수의 사후 분포의 분산이 sparse SGLMM에 비해 크다는 사실이다. GINI 계수에 해당되는 회귀계수의 증대된(inflated) 사후분포 분산 때문에 ICAR 모형에서는 GINI 계수가 유효하지 않는 것으로 보인다.

4. 결론

앞에서 살펴본바와 같이 공간적 의존관계를 회귀모형에 포함시키기 위하여 다양한 시도가 있어왔는데 특히 이산형 관측값을 모형화하는 데에 있어서는 계층적 구조로 모형화하여 잠재(latent) 계층에 공간 프로세스를 도입하고 공간 프로세스가 주어진 단계에서는 관측값간의 독립을 가정하는 형태의 혼합모형이 일반적으로 사용되었다. 특히 일반화된 선형 혼합모형(SGLMM)과 빅데이터를 다루기 위한 여러 기술들을 알아보았고 R 페키지를 활용하여 실제 데이터에 SGLMM을 적합하여 활용방법을 제시하였다.

References

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society B*, **70**, 825–848.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society, Series B*, **36**, 192–236.

Caragea, P. and Kaiser, M. (2009). Autologistic models with interpretable parameters, *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 281–300.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, 2nd edition, Wiley, New York.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 299–350.

Hartman, L. and Hössjer, O. (2008). Fast kriging of large data sets with Gaussian Markov random fields, *Computational Statistics and Data Analysis*, **52**, 2331–2349.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models, *Journal of the Royal Statistical Society: Series B*, **75**, 139–159.

Hughes, J. and Cui, X. (2015). ngspatial: Fitting the centered autologistic and sparse spatial generalized linear mixed models for areal data. R package.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models, *Applied Statistics*, **52**, 1–18.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000). Smoothing spline ANOVA models for large datasets with Bernoulli observations and the randomized GACV, *The Annals of Statistics*, **28**, 1570–1600.

Lindgren, F., Lindström, J. and Rue, H. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach, *Journal of the Royal Statistical Society B*, **73**, 423–498.

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large datasets, *Computational Statistical Data Analysis*, **51**, 3631–3653.

Reich, B., Hodges, J. and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models, *Biometrics*, **62**, 1197–1206.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC, Boca Raton.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society B*, **71**, 319–392.

Rue, H., Martino, S., Finn, L., Simpson, D., Riebler, A. and Krainski, E. T. (2014). INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using integrated nested Laplace approximation. R package, version 0.0-1404466478.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian field, *Scandinavian Journal of Statistics*, **29**, 31–49.

Wikle, C. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering, *Biometrika*, **86**, 815–829.

공간적 상관관계가 존재하는 이산형 자료를 위한 일반화된 공간선형 모형 개관

박진철[a,1]

[a]계명대학교 통계학과

(2015년 4월 1일 접수, 2015년 4월 6일 수정, 2015년 4월 6일 채택)

요 약

공간적으로 관측되는 연속형 자료를 분석하는 모형으로 공간적 상관관계를 고려한 다양한 정규모형이 지난 수십 년간 제안되었다. 그 중에서 공간효과를 랜덤효과로 모형화하는 공간선형모형(Spatial Linear Mixed Model; SLMM)이 가장 널리 활용되는 모형 중 하나일 것이다. 연결함수(link function)을 사용하면 SLMM을 비정규 데이터도 적용할 수 있는 일반화된 공간선형모형(Spatial Generalized Linear Mixed Model; SGLMM)으로 자연스럽게 확장할 수 있다. 이 논문에서는 가장 널리 활용되는 SGLMM을 알아보고 실제 데이터 적용사례를 R 패키지를 활용하여 제시하고자 한다.

주요용어: 비정규 데이터, 일반화된 공간선형모형.

[1](704-701) 대구광역시 달서구 달구벌대로 2800, 계명대학교 통계학과. E-mail: park.jincheol@gw.kmu.ac.kr