응용통계연구 제 14권 2호, 2001년, pp. 379-391

확률난수를 이용한 공간자료의 생성과 베이지안 분석*

이윤동[1)]

요 약

본 연구에서는 관심거리가 되고 있는 마코프연쇄 몬테칼로(Markov Chain Monte Carlo, MCMC) 방법에 근거한 공간 확률난수(spatial random variate) 생성법과 깁스표본추출법(Gibbs sampling) 에 의한 베이지안 분석 방법에 대한 기술적 사항들에 관하여 검토하였다. 먼저 기본적인 확률난수 생성법과 관련된 사항을 살펴보고, 다음으로 조건부명시법(conditional specification)을 이용한 공간 확률난수 생성법을 예를 들어 살펴보기로 한다. 다음으로는 이렇게 생성된 공간자료를 분석하기 위하여 깁스표본추출법을 이용한 베이지안 사후분포를 구하는 방법을 살펴보았다

주요용어: 공간 확률난수, 조건부 명시법, 메트로폴리스-해스팅 알고리즘.

1. 서론

시계열 자료는 시간지수에 대하여 자기상관을 갖는 자료를 말하는데 대하여, 공간자료의 경우는 공간지수에 대한 자기상관, 다시 말하여, 공간상관(spatial correlation)이라는 특성을 갖는 자료들이다. 이 분야는 최근 기술 발달의 영향으로 점점 더 많은 관심의 대상이 되고 있다. 공간자료에 대한 연구는, 모든 통계학 분야가 마찬가지이겠으나, 다른 어떤 분야보다도 자료 중심의 연구가 꼭 필요한 분야라고 알려져 있다. 그럼에도 불구하고 보통 공간자료의 수집과 처리에는 많은 시간과 비용이 소요되어서, 공간자료 해석과 관련된 이론 연구에 관심 있는 사람들의 경우에 자신의 이론이 대상으로 하는 마땅한 자료를 찾아 이를 시험해 보기 어려운 점이 있다. 이러한 문제를 해결하기 위한 방법으로 사용되는 것이 가상적으로 자료를 생성해내는 공간 확률난수를 이용한 시뮬레이션 방법이다. 이런 시뮬레이션 방법은 공간자료의 문제와는 별개로 통계학 연구의 중요 부분으로 연구되어져 왔다. 특히 최근에는 베이지안 통계학의 한 중요한 방법론으로서 그 의미를 더하고 있는 듯 하다. 본 연구에서는 공간종속성을 가진 자료를 만들기 위해 필요한 확률 난수 생성법을 살펴보고, 이렇게 생성된 공간 확률난수들을, 다시 확률난수를 이용한 베이지안 분석법에 따라 분석해 보는 과정을 보임으로써, 공간자료의 가상적 생성과 그 분석 과정에 공통적으로 확률난수 생성법이 응용되는 것을 대비적으로 살펴볼 수 있는 기회를 제공하고자 한다. 먼저 공간 확률난수 생성법에 대하여 거론하기에 앞서 일반적인 확률난수 생성법과 관련된 사항 중에서 이후 언급될 내용과 관련된 기본 사항을 먼저 살펴보기로 한다.

* 본 연구는 서울대학교 복잡계통계 연구센터를 통한 한국과학재단의 지원에 의하여 수행되었음.

1) (151-742) 서울시 관악구 신림동, 서울대학교 복잡계통계연구센터, Reserch Fellow
E-mail: poisson@dreamwiz.com

2. 확률난수

역시 모든 시뮬레이션 관련 이론이나 방법론 연구에서 그 밑바탕을 형성하는 것은 어떻게 좋은 성질을 갖는 정수형 균일분포의 난수를 얻을 것이냐 하는 점이다. 보통 많이 사용되는 난수생성 알고리즘으로는 Lehmer(1951)계열의 선형, 비선형, 다중 방식을 사용한 다양한 합동생성기들이나, 레지스터의 이진 정보를 비트 단위로 조작하는 GFSR(Generalized Feedback Shift Resister)과 같은 Tausworth(1965)계열의 방법들이 있다. 이러한 여러 가지 난수 생성 알고리즘들은 각각 다른 속도와 주기를 가지고 있다. 속도와 주기라는 특성 외에도 이들 난수 생성기들은, 얼마나 균일분포의 특성을 잘 만족하는가 하는 점, 각각 순차적으로 생성되는 값들의 독립성과 관련하여 얼마나 잘 무예측성(ergodicity)조건을 만족하는가 하는 점, 또 실제로 구현된 코드의 간결성 이라는 기준 등등에 대하여 각기 다른 특성을 보여주고 있다.

SAS, Splus 등등의 통계 소프트웨어와, 잘 알려진 수학용 라이브러리 함수들은 보통 2.1×10^9 주기를 갖는 것으로 알려져 있다. 그러나 최근 통계학 분야에서 진행되는 연구들 가운데 규모가 큰 경우는 이 정도의 주기를 갖는 난수 생성기로는 감당하기 어려운 경우들이 있다. 그래서 특별히 작은 주기의 난수 생성기들을 결합 사용하여 큰 주기의 난수 생성기를 얻는 방법이 사용된다. L'Ecuyer(1988)가 제안한 결합 방법을 사용하는 난수 생성기 중에 2.0×10^{88} 이상의 주기를 갖는 경우들이 발표되어 사용되고 있다.

여러 가지 난수 생성기들의 균일분포성과 무예측성을 시험하기 위하여 고안된 시험 방법들은 크게 알고리즘 중심의 시험법과, 발생된 난수를 직접 검토하는 경험적 시험법으로 나눌 수 있다. Coveyou and MacPherson(1967)의 스펙트럴시험법(spectral test)과 Marsaglia (1972) 격자시험법(lattice test) 등은 알고리즘 시험법에 속하고, 다이하드 시험법(Diehard Tests)으로 통칭되는 일련의 시험법들은 경험적 시험법에 속한다. 이 다이하드 시험법은 생성된 난수들의 값을 이진 비트화 시켜서 그 이진수들 사이에 어떤 통계적 연관성이 있는지를 찾는 방식으로 이루어진다. 여기에는 birthday spacing test, DNA test, parking lot test 등의 18가지 시험 방법이 포함되어 있다. McCullough(1998)은 SAS, SPSS, Splus 등의 통계 소프트웨어가 제공하는 난수 생성기들이 이들 시험에 일부 실패하고 있음을 보여 주고 있다.

발생된 난수를 실제 통계적인 응용에 사용하기 위해서는, 난수가 다양한 형태의 분포를 갖는 확률난수가 되도록 변환하는 과정이 필요하다. 이를 위해서는 여러 가지 방법이 고안되어 있다. 그 중에서 응용범위가 가장 넓고 중요한 방법이 기각표본추출법이라 하겠다. 기각표본추출법의 가장 큰 단점은 역시 적당한 봉투함수(envelope function)를 잡기가 힘들다는 점이다. 잘못된 봉투함수 설정은 확률난수 생성기의 성능을 급격히 떨어뜨리게 된다. 밀도함수의 식 계산에 시간이 많이 걸리는 경우, 아랫봉투함수(lower envelope function)라고 불리는, 계산이 용이하고 밀도함수보다 작은 값을 갖는 함수를 설정하여 난수 생성 알고리즘의 속도를 개선시키는 방법이 사용된다. 이러한 아랫 봉투함수에 대하여, 원래의 봉투함수를 상대적으로 윗봉투함수(upper envelope function)라고 부르기도 한다. Gilks(1992)는 log-concave 분포들에 대하여 표본추출과정 중에 윗봉투함수와 아랫봉투함수가 자동적이고 점진적으로 개선이 되는 적응적 기각표본추출법(Adaptive Rejection Sampling, ARS)을

제안하였다.

3. 공간 확률난수

공간상의 점들 $S = \{s_1, \ldots, s_n\}$에 정의된 확률변수 $Z(s)$, $s \in S$ 들에 대하여는, 그들 사이의 종속성을 나타내기 위한 척도로 $2\gamma(h) = Var(Z(s+h) - Z(s))$와 같이 정의된 배리오그람(variogram)이라는 값이 사용된다. 공간 S 의 각 점 s_i 들이 시계열에서와 같이 1차원 값을 갖는다고 가정할 때, 공간 확률과정 $Z(s)$, $s \in S$ 가 시계열에서 정의된 약안정성(weak stationarity)을 만족하는 경우, 배리오그람과 시계열의 자기공분산함수 $C(h)$사이에는 이론적으로, $2\gamma(h) = 2(C(0) - C(h))$인 관계가 성립한다. 그러나 배리오그람의 모형화에는 두 가지 면에서 시계열 자기공분산함수의 모형화와 다른 점이 있다. 하나는, 조각효과(nugget effect)라고 불리는 것으로, 이론적으로 당연한 $2\gamma(0) = 0$ 이라는 조건이 절대적으로 받아들여지지는 않는다는 점이다. 다른 하나는, 시계열모형에서는 자기공분산함수 값이 항상 유한하게 정의되는 약안정성 모형을 가정하게 되나, 공간확률 모형에서는 $2\gamma(h)$ 값이 h에 따라 무한대로 커질 수 있는 기본안정성(intrinsically stationary) 모형을 상정한다는 점이다. 공간 S 의 각 점 s_i 들이 2차원 이상의 값을 갖는 경우, $2\gamma(h)$가 $2\gamma(h) = ||h||^{\theta}$ 와 같이, h 에 대하여 노옴 $||h||$ 만의 함수로 나타나게 되면, 이러한 공간종속모형을 등방향성(isotropic)이라고 하고, 그렇지 않은 경우를 비등방향성(anisotropic)이라고 한다. 공간확률변수들이 갖는 종속성은 이와 같이 배리오그람이나, 코배리오그람(covariogram)이라고 불리는 (다차원 공간상에 적용이 가능하도록 확장된 정의를 갖는) 자기공분산함수에 의해서 나타나게 되는데, 다음에서는 이와 같은 방법으로 공간종속성이 규정되는 공간 확률과정을 가상적으로 생성하기 위해 사용되는 방법을 살펴보기로 한다.

먼저 촐레스키(Cholesky) 분해법은, 다변량 정규분포 생성법을 직접 이용하는 방법이다. 각 확률변수 사이에 공간적 특성에 따라 미리 지정된 배리오그람이나 자기공분산 함수에 따라, 이에 대응하는 다변량 정규분포의 공분산 행렬을 얻고 이를 촐레스키 분해한 다음, 정규분포 확률난수 생성기를 통하여 얻어진 값을 앞서 촐레스키 분해를 통하여 얻어진 행렬을 이용해서 선형 변환함으로써 공간 확률난수를 얻는 방법이다. 그러나 이 방법은 시계열 자료 생성 경우에서 볼 수 있는 예들로부터 이미 잘 알려진 바와 마찬가지로, 매우 큰 행렬 연산이 필요하게 되고, 그 결과의 정확성이 떨어진다는 치명적인 단점을 가지고 있다.

그 외의 잘 알려진 방법으로 Shinozuka and Jan(1971)과 Mejia and Rodriguez-Iturbe (1974) 등이 제안한 스펙트럴 방법이 있다. 이 방법은 주어진 자기공분산함수로부터 이에 대응하는 스펙트럴 밀도함수를 얻고, 이 밀도함수로부터 얻어지는 iid 확률변수를 생성시켜 퓨리에 변환을 함으로써 원하는 종속성을 갖는 공간 확률변수를 얻는 방법이다. 이 방법은 앞서 말한 촐레스키 분해법을 사용하는 방법과는 달리 매우 큰 크기의 행렬을 다룰 필요가 없다는 장점을 가지고 있고, 이론적 우아함을 갖추고 있지만, 일부에서 이런 방법으로 생성된 확률난수에 무예측성 문제에 대한 결함이 있을 수도 있다는 논의가 있다(참조 Chiles and Delfiner, 1999). 그러나 무예측성 결여 문제는 특별히 하나의 공간 확률난수를 만드는데 필요한 기본 확률난수의 수를 극단적으로 줄이지 않는 한 실제적인 경우에 있어

서 큰 문제가 되어 보이지는 않는다. 대신, 이러한 스펙트럴 방법의 가장 큰 단점은, 주어진 배리오그람 또는 자기공분산함수로부터 스펙트럴 밀도함수를 얻기가 힘들다는 점이다. 여러 가지 방법을 통해서 스펙트럴 밀도함수가 얻어진다고 하더라도, 이렇게 얻어지는 밀도함수의 경우에 그 형태가 복잡하여, 그 밀도함수를 갖는 확률난수를 얻기 위해서는 기각표본추출법과 같은 다른 방법을 중복하여 사용해야만 한다는 점이다. 더욱이 문제가 되는 점은 공간지수가 2차원 이상이 되는 경우에, 일상적으로 많이 사용되는 배리오그람 모형들에 대하여서도 특수한 경우를 제외하고는 그 스펙트럴 밀도함수가 얻어지지 않는다는 점이다. 물론 이의 해결을 위해서 Mantoglou and Wilson(1982)은 회전띠(Turning-Bands)방법을 제시하고 있으나 이 방법도 적용 범위가 마찬가지로 넓지 않다는 단점을 가지고 있다. 공간종속성의 형태가 방향에 관계 없이 일정한 성질을 갖는 등방향성인 경우는 복잡한 과정을 거쳐서라도 회전띠 방법을 적용하는 것이 가능하나, 공간종속성의 형태가 비등방향성인 경우는 회전띠방법의 적용이 불가능하다.

위의 두 방법에 비하여, Besag(1974)이 제안한 조건부명시법(conditional specification)에 의한 공간 확률난수 생성법은 그 활용도가 넓어서 많은 장점을 가지고 있다. 특히 이 이론은 최근의 베이지안 이론들과 그 틀을 같이하고 있어서 많은 주목을 받고 있다. 시계열 확률난수 생성의 경우, 시계열 모형이 주어지면 과히 어렵지 않게 그 모형에 대응하는 자기공분산함수를 갖는 시계열 확률난수 생성이 가능하다.

$$Z_n = \theta Z_{n-1} + \epsilon_n, \qquad \epsilon_n \sim iid\ Normal(0, \sigma^2) \tag{3.1}$$

과 같이 주어진 AR(1) 모형의 경우, Z_{n-1} 의 값이 z_{n-1} 로 주어져 있을 때, Z_{n-1} 은 θz_{n-1} 를 평균으로 갖는 정규분포로부터 쉽게 얻을 수 있다. 이런 방법을 순차적으로 반복함으로써 모형에 맞는 확률난수 생성이 가능하다. 다시 말하여, 시계열 모형 (3.1)은

$$Z_n | Z_{n-1} \sim Normal(\theta z_{n-1},\ \sigma^2), \qquad n = 1, 2, \cdots$$

와 동일한 의미를 갖는다는 것이다. 그러나 시계열의 시간지수와 달리 공간 확률모형에서 공간지수는 보통 2차원 이상인 경우가 많고, 1차원인 경우도 순서가 정의될 만한 지수의 방향성이 없기 때문에, 앞서와 같은 성질이 성립하지 않는다. 예를 들어

$$Z(s) = \theta Z(s-1) + \theta Z(s+1) + \epsilon(s), \qquad \epsilon(s) \sim iid\ Normal(0, \sigma^2) \tag{3.2}$$

이라고 정의된 공간 확률모형의 경우,

$$Z(s) | Z(s-1), Z(s+1) \sim Normal(\theta\{z(s-1) + z(s+1)\}, \sigma^2), \quad s \in S \tag{3.3}$$

와 같은 방법으로 모형에 맞는 확률난수를 생성할 수 없다. 대신에 식 (3.3) 자체를 공간확률과정 $Z(s), s \in S$를 정의하기 위한 모형으로 사용하게 되는데, 이 방법이 조건부명시법이다. 이에 반하여 (3.2)와 같이 모형을 설정하는 방법을 동시명시법(simultaneous specification)이라고 한다. 시계열에서는 동시명시법과 조건부명시법이 위의 경우와 같이 일치하게 되는데 반하여, 공간확률 모형에서는 동시명시법과 조건부명시법이 서로 다른 확률과정을 의미하

게된다. 조건부명시법은 동시명시법에 비하여 포괄하는 모형의 수리적 범위가 넓고, 직관적 모형화 과정이 용이하다. 이 조건부명시법은 정규분포 이 외의 모형에도 적용하는 것이 가능한데, 오토-로지스틱(auto-logistic) 모형, 오토-포아송(auto-poisson) 모형 등의 다양한 적용이 가능하다. 조건부명시법에 의한 공간 확률난수 생성법은 그 유연성과 응용성에 있어서 훌륭한 장점을 가지지만, 반면 계산량이 많고 계산 시간이 오래 걸리는 단점을 가지고 있다. 또, 매우 주기가 길고, 속도가 빠르고 신뢰성이 높은 난수 생성기를 요구한다는 점이 문제가 될 수 있다. 100개 지점에서 관측된 공간자료를 생성하려하는 경우를 가정해보자. 또 가정되는 분포가 정규분포가 아닌 경우로 기각표본추출법을 써서 분포를 생성해야 한다면, 여기서 매번 최대 100개 정도의 난수를 예비해야 하고, 이와 같은 과정을 전체적으로 10000회에서 100000회 정도 반복해야 한다고 한다면, 이미 보통의 통계 소프트웨어들이나 수치해석용 라이브러리 함수들이 제공하는 정도의 2.1×10^9주기를 갖는 난수 생성기로는 그 결과를 보장받을 수 없는 정도에 이르게 된다. 이런 경우에 대한 대안으로 인터넷을 통하여 쉽게 얻을 수 있는 randlib 과 같은 전문 라이브러리 함수의 사용이 필요하다고 보인다.

조건부명시법을 통하여 정의된 공간 확률변수 모형을 만족하는 난수 생성법을 살펴 보기로 하자. 식 (3.3)의 평균항을 좀더 일반화하여 다른 방법으로 나타내면

$$E(Z(s)|Z(t), t \neq s) = \mu_s + \sum_{t \neq s} \theta_{s,t}(Z(t) - \mu_t), \qquad s, t \in S = \{s_1, \cdots, s_n\} \tag{3.4}$$

와 같이 쓸수 있다. 이렇게 정의된 $Z = (Z(s_1), \cdots, Z(s_n))'$을 다변량 정규분포로 나타내면

$$Z \sim Normal(\mu, \sigma^2(I - C)^{-1}) \tag{3.5}$$

라고 쓸수 있는데, 여기서 $\mu = (\mu(s_1), \cdots, \mu(s_n))'$이고, C는 대각원소는 0이고 s행과 t열에 $\theta_{s,t}$값을 갖는 크기 $n \times n$행렬이다. 여기서 식 (3.3)의 분산항을 비등분산을 갖는 것으로 일반화하는 경우에, 식 (3.5)의 분산항은, σ^2 대신 비등분산성을 갖는 분산항들을 그 대각원소로 갖는 행렬 M에 대하여 $(I - C)^{-1}M$와 같은 방법으로 나타나게 된다. Z의 각 값들에 대하여 임의의 초기값을 부여하고 식 (3.2)의 방법을 반복적으로 여러 번 적용함으로써 식 (3.5)에 해당하는 확률변수를 얻는 방법이 조건부명시법을 이용한 공간 확률난수 생성법이다. 위의 식 (3.2)와 (3.5)의 동일성이 바로 조건부명시법이 촐레스키 분해법을 이용한 경우에서와 달리 크기가 큰 행렬을 다루지 않고서도 공간 확률난수를 생성시킬 수 있는 이유이다. 조건부명시법에서는 $\theta_{s,t}$를 어떻게 정의하느냐에 따라 다양한 형태의 공간종속성을 정의할 수 있다. 다음에서는 가장 단순한 형태의 공간종속성을 갖는 경우를 예로 들어보기로 한다.

표 3.1은 $n = 10$개의 관측점에 대한 평면 위치 정보 $s = (x, y)$에 관한 자료와, 다음에서 설명하는 값을 주고 조건부명시법을 이용하여 각 관측점에서 얻은 공간 확률난수 $Z(s)$들이다. 그림 3.1은 이 각 점들에서 얻어진 공간 확률난수를 (x, y)좌표와 함께 3차원으로 표현한 것이다. 3.1에 제시된 공간 확률난수 Z는 식 (3.5)에서 $\mu = (\mu_0, \cdots, \mu_0)'$이고

$\theta_{s,t} = \eta/dist(s,t)$라하고 $\mu_0 = 10$, $\eta = 0.2$, $\sigma^2 = 1.0$ 인 경우에 (3.2)를 반복적으로 적용하여 얻은 확률난수이다. 여기서 두 점 s와 t의 관측 좌표 (x_s, y_s), (x_t, y_t)에 대하여 두 점 사이의 거리 $dist(s,t)$는 $|x_s - x_t| + |y_s - y_t|$로 주어졌다. 이와 같은 관계에서 식 (3.5)의 행렬 C는 공간종속성을 의미하는 모수 η의 함수가 되어 $C(\eta)$라고 표시되고, 행렬 $H = C(\eta)/\eta$는 관측점들의 좌표 (x,y)값들로부터 직접 값이 정해지는 행렬이 된다. 모수 η가 취할 수 있는 값의 범위는 행렬 $(I-C)^{-1}$가 양정치(positive definite) 행렬이 되어야 하므로 행렬 H가 갖는 고유값 $h_1 \leq \cdots \leq 0 \leq \cdots \leq h_{10}$ 들의 특성에 의하여 $h_1^{-1} < \eta < h_{10}^{-1}$ 로 제약되게 된다. 표 3.1의 (x,y) 좌표들에 대하여 $h_1^{-1} = -0.29983$, $h_{10}^{-1} = 0.249759$로 주어진다.

표 3.1: 좌표 (x,y)에 대한 Z(x,y)의 값

x	y	$Z(x,y)$
3.31	2.87	11.616
3.18	3.04	12.204
4.32	2.34	10.130
4.69	3.86	9.4883
6.31	3.21	10.061
7.97	3.43	9.8264
9.12	5.79	8.7610
4.17	5.76	10.509
2.59	7.32	9.8935
3.26	9.18	7.6811

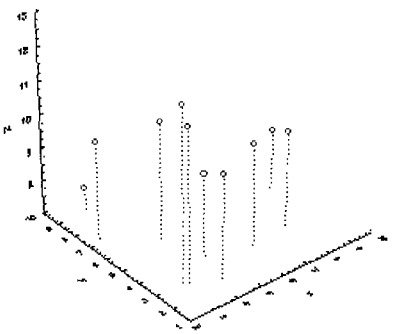

그림 3.1: 그래프로 나타낸 좌표 (x,y)에 대한 Z(x,y)의 값

4. MCMC 방법과 베이지안 분석

MCMC 방법의 범주에는 담근모사법(simulated-annealing)과 같은 확률적 함수 최적화(stochastic optimization)문제도 포함된다고 할 수 있으나, MCMC 방법의 주된 응용 분야는 복잡한 형태의 분포로부터 난수를 생성시키는 일반적인 난수 생성 기능이다. MCMC 방법의 원류인 메트로폴리스-해스팅(Metropolis-Hasting, MH) 알고리즘은, Metropolis et.al. (1953)이 이산상태공간(discrete state-space) 모형에서 함수의 최적화 문제를 다루면서 제시한 알고리즘을, 후에 Hasting(1970) 등이 개선하여 얻어진 것이다. 이후 Geman and Geman(1984)이 이미지 분석 문제에 응용하면서, 통계적 응용이 보편화되었다. MH 알고리즘은 담근모사법 외에도 입자물리학에서 사용되는 볼쯔만(Boltzman) 알고리즘등과 같은 유사한 개념들의 특수한 경우로 이해되어지기도 한다. MH 알고리즘은 목적분포(target distribution) $\pi(\cdot)$를 따르는 확률난수 X를 얻기 위해, 임의의 초기값 X_0에 대하여 $X_\infty \sim \pi(\cdot)$인 성질을 갖게되는 확률변수열 $\{X_i\}_0^\infty$를 생성하는 방법이다. 보통은 적당히 큰 수 N에 대하여 $\{X_i\}_0^N$을 생성시키고 X_N을 X에 할당하는 방법을 사용한다. 이 알고리즘의 핵심적 사항은 확률변수열 $(X_0, \cdots, X_i)$의 다음 값 X_{i+1}을 선택하는 방법이다. 이를 위해서 먼저 좋은 성질을 갖는 제안함수(proposal function) $q(\cdot|X_i)$와 이에 대응하는 밀도함수로부터 확률변수 Y를 얻고, 변환허용비율

$$\alpha(X_i, Y) = min(1, \pi(X_i)q(Y|X_i)/\pi(Y)q(X_i|Y))$$

값에 따라, $\alpha(X_i, Y)$만큼의 확률로 $X_{i+1} = X_i$라고 하는 방법이 사용된다. MH 알고리즘은, 진행 시점의 분포가 목적분포를 따른다고 가정했을 때, 현재 상태 X_i로부터 제안 상태 Y로의 변환확률 $\pi(X_i)q(Y|X_i)$와, Y로부터 X_i로의 변환확률 $\pi(Y)q(X_i|Y)$이 동등하게 되도록 변환허용비율 $\alpha(X_i, Y)$를 이용하여 인위적으로 조정하여 줌으로써, 목적분포의 상태를 항상 유지시켜주도록 하고 있다. 나아가서 진행시점에서의 분포가 목적분포가 아닌 경우에는 점진적인 과정을 거쳐서 결국은 안정 상태인 목적 분포에 이르도록 고안된 알고리즘이다. MH 알고리즘은 기각표본추출법, 특히 ARS 방법과 매우 유사한 구조를 가지고 있음을 볼 수 있다. Gilks et. al.(1995)이 제안한 ARMS(Adaptive Rejection Metropolis Sampling) 방법은, MH의 제안함수가 점진적 과정을 거쳐 ARS의 봉투함수 역할을 할 수 있도록 함으로써, log-concave 조건을 만족하지 않는 밀도함수를 가진 경우에도 적용할 수 있도록 ARS 방법을 개선한 것이다. MH 알고리즘과 유사성을 가진 담근모사법은 변환허용비율 $\alpha(X_i, Y)$를, 최적화 하고자 하는 목적함수 $g(\cdot)$의 X_i와 Y의 값 $g(X_i)$와 $g(Y)$에 대한 함수로, 예를들어 $exp(g(Y) - g(X_i))$와 같이 알맞게 정의함으로써, 확률적 과정을 통해 최종적으로 함수 $g(\cdot)$의 전역적 최적값(global optimal value)을 찾고자 하는 방법이다.

MH 알고리즘은 적용되는 곳에 따라 고차원 분포로부터 확률난수를 생성시키기 위한 깁스표본 추출법이나, $q(X_i, Y) = q(Y - X_i)$인 형태의 제안함수를 이용하는 무작위보행표본추출법(random work sampling), $q(X_i, Y) = q(Y)$인 형태의 성질을 만족하는 제안함수를 이용하는 독립표본추출법(independent samppling)등이 있다. 공간 확률난수 생성을 위한 식 (3.3)는 깁스표본추출법의 한 응용이라고 볼 수 있다. 특히 깁스표본추출법은 복잡한 통계모형에 대한 베이지안 추론을 위하여 많이 사용되고 있다. 베이지안 추론의 사후분포는

모형에 사용된 여러 개의 모수가 동시에 개입되어 있고 각각이 서로 다른 분포를 갖고 있어서, 이들의 결합분포를 한 번에 처리하기 어려운 점이 있다. 대신 하나의 모수만을 제외하고 다른 모수들의 값을 조건으로 주어진 것으로 가정하고 얻어내는, 사후 조건부 주변분포를 이용하여 반복적인 깁스표본추출법을 통해서 전체 모수들에 대한 사후분포를 얻게된다. 이와 같이 주어진 사후 조건부 주변분포를 전조건분포(full conditional distribution)라고 한다.

앞 절에서 얻은 공간자료를 분석하는데 있어서 식 (3.5)를 우도함수로 하고 μ_0에 대하여는 균일 사전분포(flat prior)를 가정하였고, $(\sigma^2)^{-1}$에 대하여는 (사전분포의 영향이 최소화 되도록) 형태모수가 0.5이고 규모모수가 2인 감마분포를 가정하였다. 이 모형의 특이점은 공간종속성을 나타내는 모수 η가 개입되어 있다는 점인데 η에 대하여는 (h_1^{-1}, h_{10}^{-1}) 범위 내에서 균일한 값을 갖는 균일 사전분포를 가정하였다. 공간종속 모수 η에 대한 사전분포를 일반적으로 가정하기 위해서는, 행렬 H의 고유치들이 어떻게 변하느냐에 따라서 η가 취할 수 있는 값의 범위나 구간의 형태가 달라진다는 점을 고려해야 한다. 구간의 형태에 따라 베타 사전분포나 감마 사전분포를 일반적인 형태로 가정하는 것도 가능하리라 보인다. 분산항 σ^2에 대하여는 보통 역감마 분포나 로그노말 사전분포를 가정하는 것이 가능하다. Daniels et. al.(2001)은 대기중 분진의 양을 해석하기 위한 모형에서 등분산성의 검토를 위해서 분산항에 대하여 로그노말 사전분포를 가정하였다.

이러한 사전분포들을 가정할 때, 사후분포로부터 얻어지는 μ_0에 대한 전조건분포는 다음과 같다.

$$\mu_0|Z,\sigma,\eta \ \sim \ Normal(m, s^2)$$

여기서 $u=(1,\cdots,1)'$일 때 $m=(Z'(I-\eta H)u)/(u'(I-\eta H)u)$이고 $s^2=(u'(I-\eta H)u)^{-1}$이다. σ^2에 대한 전조건분포는 다음가 같이 주어진다.

$$(\sigma^2)^{-1}|Z,\mu,\eta \ \sim \ Gamma(5.5, b),$$

여기서 $b=2/(1+(Z-\mu)'(I-\eta H)(Z-\mu))$이고 $\mu=(\mu_0,\cdots,\mu_0)'$이다. η에 대한 전조건분포 $f(\eta|Z,\mu,\sigma)$는 어떤 잘 알려진 형태의 분포로 나타나지 않고 다음과 같이 표현된다.

$$f(\eta|Z,\mu,\sigma) \ \propto \ \prod_{i=1}^{10}(1-\eta h_i)^{0.5}\cdot exp(-0.5(Z-\mu)'(I-\eta H)(Z-\mu))$$

위와 같은 형태로 주어진 η의 전조건 분포로부터 표본추출을 위해서는, 역시 MH 방법중의 하나인 독립표본추출법이 사용될 수 있다. 독립표본추출법을 사용하는 경우에 제안함수로 (h_1^{-1}, h_{10}^{-1}) 사이에 정의된 균일분포를 사용하면, 변환허용비율 $\alpha(\eta_i, \eta_{i+1})$는

$$f(\eta)=\prod_{i=1}^{10}(1-\eta h_i)^{0.5}\cdot exp(\eta(Z-\mu)'H(Z-\mu)/2)$$

라고 할 때 $\alpha(\eta_i,\eta_{i+1})=\beta(\eta_{i+1})/\beta(\eta_i)$가 된다.

그림 4.1, 그림 4.2, 그림 4.3는 이와 같은 방법으로 구해진 (μ,σ^2,η)의 사후분포에 대한 주변분포를 알아보기 위해서 5000개의 표본을 이용해서 그린 히스토그램들이다. 이들

5000개의 표본은, 깁스표본추출법을 사용하여 20000회의 번인(burn-in) 과정 후에, 표본들 사이에 독립성을 유지시키기 위해서 각 50회의 반복을 간격으로 두고 뽑은 것이다. 앞 절에서 공간 확률난수 생성법을 이용해서 자료 Z를 구할 때 $\mu = 10, \sigma^2 = 1.0, \eta = 0.5$인 값을 사용했었다는 점과 일관된 결과를 각 히스토그램들로부터 알아볼 수 있다. 그림 4.1은 평균이 대략 10근처에 형성되어 있고 정규분포에 가까운 모습을 보여주고 있고, 그림 4.2은 평균이 대략 1정도가 되는 감마 분포의 형태를 보이고 있다. 또한 위의 깁스표본추출 과정에 대하여 다양한 초기값을 이용한 반복 실험에서도 안정적이고 일관적으로 주어진 그림들과 거의 동일한 결과를 얻을 수 있다.

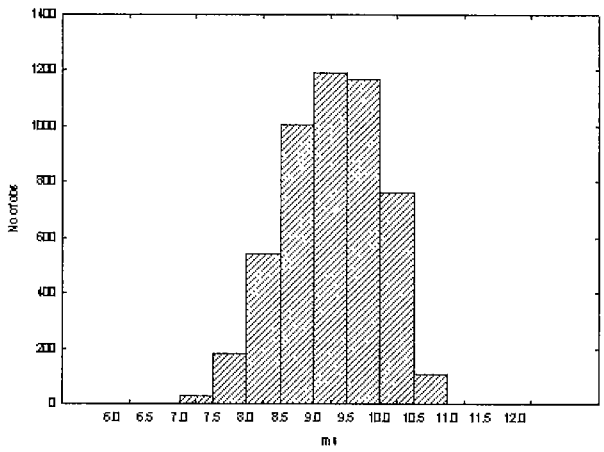

그림 4.1: μ의 사후 주변 분포

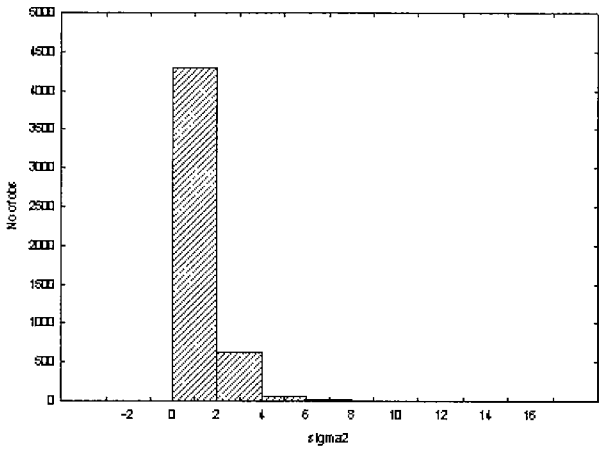

그림 4.2: σ의 사후 주변 분포

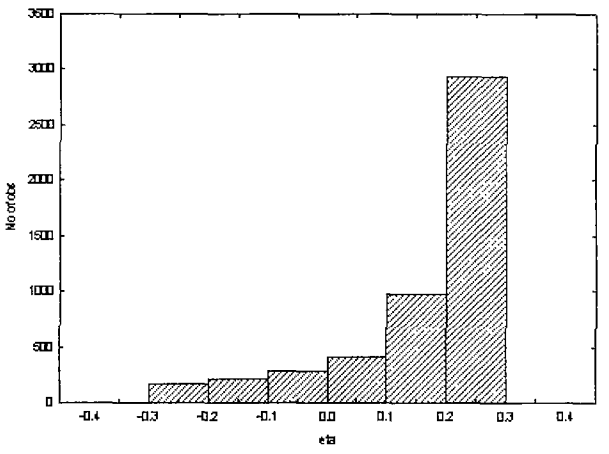

그림 4.3: η의 사후 주변 분포

5. 맺음말

앞 절들에서 우리는 공간 확률난수 생성에 필요한 제반 기술적 사항을 검토하였다. 촐레스키 방법, 스펙트럴 방법, 조건부명시법 등등 공간 확률난수 생성 방법을 알아보았고, 각 방법들이 가지는 장단점에 대하여 살펴보았다. 또한 이렇게 생성된 자료에 대하여 베이지안 깁스표본추출법을 이용한 베이지안 추론법을 적용함으로써 처음 생성 방식과 일관성을 갖는 결과를 얻을 수 있음을 보였다. 위의 과정을 통하여 Besag(1974)이 제안한 공간확률 과정의 조건부 명시법에 의한 정의와, 최근 활발한 이론적 연구가 진행되었던 MCMC 방법이 이론적으로 동일한 틀 안에서 해석되어 질 수 있음을 보다 직접적이고 대칭적인 관계로 설명하였다.

감사의 글

본 연구 수행 기간동안 도움을 준 StatSoft의 Dr. Shin과, 좋은 조언을 해 주신 심사위원들께 감사 드립니다.

참고문헌

[1] Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice system, *Journal of the Royal Statistical Society*, B, **36**, 192-225.

[2] Chiles, J. and Delfiner, P. (1999).*Geostatistics*, Wiley, New York.

[3] Coveyou, R.R. and MacPherson, R.D. (1967). Fourier analysis of uniform random number generators, *Journal of the ACM*, **14**, 100-119.

[4] Cressie, N. (1993). *Statistics for Spatial Data*, Wiley, New York.

[5] Daniels, M., Lee, Y. and Kaiser, M. (2001). Assessing sources of variablility in measurement of ambient particulate matters, *Environmetrics*, will appear.

[6] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Mach. Intelll.*, **6**, 721-741.

[7] Gilks, W.R. (1992). *Derivative-free adaptive rejection sampling for Gibbs sampling*, In Bayesian Statistics 4, pp 641-649, Oxford University Press, Oxford.

[8] Hasting, W.K. (1970). Monte Carlo Sampling methods using Markov chains and their application. *Biometrika*, **57**, 97-109.

[9] L'Ecuyer, P. (1988), Efficient and portable combined random number generator, *Communications of the ACM*, vol. **31**, pp 742-747

[10] Lehmer, D.H. (1951). Mathematical methods in large-scale computing units, *Proceedings of the second the second Symposium on Large Scale Digital Computing Machinery*, Harvard University Press, Cambridge, Massachusetts. 141-146.

[11] Marsaglia, G. (1972). *The structure of linear congruential sequences*, Applications of Number Theory to Numerical Analysis, Academic Press, New York, 294-286.

[12] McCullough, B.D. (1998), Assessing the Reliability of Statistical Software, Part II, *The American Statistician*, vol. **53**, pp149-159.

[13] Mejia, J.M. and Rodriguez-Iturbe, I. (1974). On the synthesis of random field sampling from the spectrum: An application to the generation of hydrologic spatial processes, *Water Resources Research*, **10**, 705-711.

[14] Mantoglou, A. and Wilson, J.L. (1982). The turning bands method for simulation of random fields using line generation by a spectral method, *Water Resources Research*, **18**, 1379-1384

[15] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equa-

tions of state calculations by fast computing machines, *J. of Chem. Phys.*, **21**, 1087-1092

[16] Shinozuka, M. and Jan, C.M. (1972). Digital simulation of random processes and its applications, *Journal of Sound and Vibration*, **25**, 111-128

[17] Tausworthe, R.C. (1965). Random numbers generated by linear recurrence modulo two, *Mathematics of Computation*, **19**, 201-209.

[2000년 12월 접수, 2001년 8월 채택]

Computing Methods for Generating Spatial Random Variable and Analyzing Bayesian Model

YoonDong Lee[1)]

ABSTRACT

The theoretical analogy between conditional model specification (CMS) for spatial random field and Gibb's sampling mechanism was plausibly demonstrated in a practically applicable example. In the example, the data were generated by CMS, and analyzed with Gibb's sampling method for the Bayesian model fitted to the data. Other important methods used in generating spatial random field were also briefly reviewed. A method to incorporate spatial dependence in Bayesian modeling for spatial data was tentatively tried.

Keywords: Spatial random variate; Conditional model specification; Metropolis-Hasting algorithm.

1) Research Fellow, Statistical Research Center for Complex System, Seoul National University. E-mail: poisson@dreamwiz.com