응용통계연구 제 **13**권 1호, 2000년, *pp. 189–196*

포아송-로그정규분포 모형에 관한 연구

김용철[1)]

요 약

포아송 분포에서 일반적으로 공액 사전분포를 이용하여 사후확률의 수학적 계산이 간편하도록 한다. 그러나 모수 집합의 제한적 조건 때문에 비공액 사전 분포를 이용할 수 도 있다. 비공액 사전분포의 사용은 사후분포의 형태가 일상적인 분포 집합의 형태를 갖지 않으므로 모형의 가정에 따라서 복잡한 구조를 갖을 수 도 있다. 특히 포아송-로그정규분포 모형에서의 모수 추정문제를 몬테 칼로방법을 이용하여 추정하고자 할 때 필요한 완전한 조건부 분포의 형태는 잘 알려진 분포의 형태를 갖지 않는다. 본 논문에서는 계층적 구조를 갖는 포아송- 로그정규분포 모형에 대하여 고찰하고 추정에 있어서 잠재적 변수를 활용하여 필요한 난수발생이 쉽도록 하는 방법에 대하여 알아보았다.

주요용어: 포아송-로그정규분포 모형, 몬테칼로방법, 잠재적변수.

1. 서론

일반적으로 공액 사전 분포의 사용은 베이지안 분석 관점에서 보면 사후분포 형태, 모수 추정에서의 호환성 그리고 수학적 계산의 이점을 가지고 사용을 하여왔다. 우도함수 $p(y|\theta)$에 대하여 사전분포함수 $p(\theta)$의 선택은 분포함수

$$p(y) = \int p(y|\theta)p(\theta)d\theta \tag{1.1}$$

의 값을 해석적으로 구하는 데에 있어서 중요한 문제점이었다. 적분의 형태가 단변량인 경우 또는 가능한 차원의 다변량인 경우에는 중요표본(importance sampling)을 이용하여 적분의 문제를 해결 할 수 있다. 만약 공액 사전분포로 $p(\theta)$를 선택하여 $p(y|\theta)p(\theta)$의 형태가 잘 알려진 분포함수의 형태를 이루게 된다면 적분의 계산은 더욱 간편하게 계산이 가능하다. 특히 포아송 분포의 모형에서는 공액 사전분포로 감마분포를 이용하면 사후분포가 감마분포의 형태를 가지므로 사후분포를 이용하여 기대값이나 신뢰구간을 구할 수 있다. 포아송-감마 분포의 모형은 사망률 분석에 대하여 Manton, Woodbury and Stallard(1981)에 의하여 제시되었다. 또한 Tsutakawa(1988)는 계층적 혼합선형모형으로써 포아송 모형을 채택하여 암 사망률을 추정하였다. 공액 사전분포의 사용의 장점은 사후분포가 잘 알려진 분포, 즉, 정규분포, 감마분포, 베타분포중 한 형태로 이루어진다면 모수 추정이 간편하기 때문에 사용한다. 그러나 모수의 제한적 조건 때문에 공액 사전분포를 사용하지 못하고 비

1) (449-714) 경기도 용인시 삼가동 470, 용인대학교 컴퓨터정보학부 조교수
E-mail: yckim@eve.yongin.ac.kr

공액 사전분포를 이용하게 되면 사후분포의 형태가 잘 알려진 분포의 형태를 갖지 못하는 경우가 종종 발생한다. 비공액 사전분포를 활용한 사후분포의 형태가 복잡한 경우 최근에 제시된 몬테칼로의 한 방법인 깁스표본(Gibbs sampling)의 활용이 가능할 것이다. 그러나 깁스표본은 각각의 모수에 대하여 완전한 조건부 분포(full conditional distribution)가 필요하며 각 분포로부터 난수의 발생이 용이하여야 한다. 난수 발생이 쉽지 않은 경우에도 변수변환이나 기술적 난수 발생 방법인 기각표본(rejection sampling) 방법을 이용하여 사후분포를 구할 수 있다.

본 논문에서는 계층적 구조를 갖는 포아송 모형에서 비공액 사전분포의 가정이 필요한 경우에 비공액 사전분포로써 로그정규분포를 활용한 모형에 대하여 논의하였다. 모형의 모수 추정에 있어서 조건부 분포의 형태가 잘 알려진 분포이면 쉽게 난수 발생이 가능하고 이를 이용하여 몬테칼로 방법으로 관심있는 모수에 대하여 추정할 수 있다. 그러나 일반적으로 비공액 사전분포를 이용한 조건부 확률분포는 복잡한 함수 식으로 표현된다. 이러한 경우에 기각표본(rejection sampling) 방법을 이용하여 사후분포를 구할 수 있지만 쉽지는 않다. 다른 방법으로 잠재적 변수를 활용하여 깁스표본에 적용이 가능한 방법이 Damien, Wakefield and Walker(1999)에 의하여 제시되었다. 본 논문에서는 효율성을 갖고 필요한 난수를 발생할 수 있는 조건부 분포를 얻는 과정에서 잠재적(latent) 변수를 도입한 조건부 분포를 유도하였다. 다음 장에서는 포아송-로그정규분포 모형에 관하여 논의하고, 제 3장에서는 추정에 필요한 완전한 조건부 분포를 유도하였다. 그리고 제 4장에서는 완전한 조건부 분포에서 흔히 발생되어지는 복잡한 형태의 분포 함수를 잠재적 변수를 이용하여 필요한 난수발생을 가 능하도록 하는 방법을 적용하여 새로운 조건부 분포를 계산하였다. 제 5장에서는 관련된 예와 결론을 논의하였다.

2. 포아송-로그정규분포 모형

독립적인 관찰치 y_i는 i 번째 분류에 속하며 평균이 $exp(\lambda_i)$인 포아송 분포를 따른다고 가정하자. 또한 $y = (y_1, y_2, \cdots, y_n)$는 n개의 분류에 관찰된 벡터이다. 계층적 모형에서 λ_i의 사전 분포 함수가 필요하다. 만약에 평균이 λ_i이면 일반적으로 포아송 분포의 평균값이 양의 값을 가져야 하므로 지수변환을 하지 않고 사용한다면 공액 사전분포로써 감마분포를 이용 할 수 있다. 그러나 가정에 손실 없이 임의 값 λ_i는 평균 μ이고 분산은 σ^2인 정규분포를 갖는다고 하면 $exp(\lambda_i)$는 로그정규분포를 따른다. 로그정규분포는 포아송 분포에서 비공액 사전분포이다. 평균 μ과 분산 σ^2에 대해서는 Jeffreys의 사전 분포를 따른다고 가정하자.

모수 $(\lambda_1, \lambda_2, \cdots, \lambda_n, \mu, \sigma^2)$의 결합 밀도함수는 다음과 같다.

$$\prod_{i=1}^{n} p(\lambda_i, \mu, \sigma^2 | y_i) \propto \prod_{i=1}^{n} f(y_i|\lambda_i) g(\lambda_i|\mu, \sigma^2) h(\mu, \sigma^2), \tag{2.1}$$

여기서,

$$f(y_i|\lambda_i) = \frac{e^{-exp(\lambda_i)}exp(\lambda_i)^{y_i}}{y_i!}, \tag{2.2}$$

$$g(\lambda_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(\lambda_i-\mu)^2}{2\sigma^2}\right\}, \tag{2.3}$$

$$h(\mu,\sigma^2) \propto (\sigma^2)^{-1}, \tag{2.4}$$

이고 $y_i = 0, 1, \cdots$, 그리고 - $\infty < \lambda_i < \infty$ 값을 갖는다.

일반적인 계층모형에서 첫 번째 계층에서의 함수적 관계식 $\phi(\lambda_i)$가 ϕ의 연결(link) 함수의 형태에 따라서 선형회귀직선의 모형을 갖는 경우도 고려될 수 있다. 이러한 일반적인 모형에 대해서는 Albert와 Chib(1997)에 의하여 제안되었다. 그러나 제한적인 모수의 조건이 주어진다면 함수적 변환을 필요로 한다. 일반적으로 포아송 모형에서는 사전분포 λ_i의 조건 형태에 따라서 공액 또는 비공액 사전분포의 사용이 가능하다. 그러나 $E(y_i|\lambda_i)$의 값은 항상 양수 값을 가져야 하므로 본 논문에서는 지수변환을 선택하였다. 모수의 추정은 Markov Chain Monte Carlo(MCMC)방법을 적용하였다. 다음 장에서는 MCMC 방법을 적용하기에 필요한 완전한 조건부 분포를 유도할 것이다.

3. 포아송-로그정규분포 모형에서 필요한 완전한 조건부 분포

포아송-로그정규분포 모형에서 MCMC 방법을 적용하기 위해 모수 (λ_1, λ_2, $\cdots$, λ_n, μ, σ^2)의 조건부 분포를 계층적 구조와 각각의 분포의 독립성에 의하여 구하면 쉽게 유도 할 수 있다.

먼저, 초모수(hyperparameter)인 μ와 σ^2은 다음과 같다.

$$p(\mu|y,\sigma^2,\lambda_{i;i=1,2,\cdots,n}) \sim N\left(\frac{\sum_{i=1}^n \lambda_i}{n}, \frac{\sigma^2}{n}\right) \tag{3.1}$$

이고

$$p(\sigma^2|y,\mu,\lambda_{i;i=1,2,\cdots,n}) \sim IG\left(\frac{n}{2}, \frac{\sum_{i=1}^n (\lambda_i-\mu)^2}{2}\right) \tag{3.2}$$

이다. 식(3.2)에서 $IG(\alpha,\beta)$는 형태(shape) 모수 α와 비율(scale) 모수 β를 갖는 역감마 분포 함수이다.

모수 μ와 σ^2이 주어졌을 때 $\lambda_1, \lambda_2, \cdots, \lambda_n$들의 독립적인 사후분포함수는 다음 밀도 함수의 곱에 비례한다.

$$\prod_{i=1}^{n} f(y_i|\lambda_i)g(\lambda_i|\mu,\sigma^2). \tag{3.3}$$

앞에서 표현한 각각의 조건부 분포에서의 난수 발생이 용이하다면 MCMC 방법을 적용하여 $(\lambda_1,\lambda_2,\cdots,\lambda_n,\mu,\sigma^2)$의 결합 확률밀도 함수로부터 λ_i의 주변확률분포를 계산할 수 있다. 모수 λ_i의 추정은 λ_i의 주변 확률분포에서 기대값을 계산함으로써 추정이 가능하다. 그러나 λ_i의 조건부 분포함수의 형태는 함수 g의 형태에 따라서 다르게 나타난다. 난수발생이 쉬운 함수의 형태를 가질 수 도 있고 그렇지 않을 수 도 있다. 특히 함수 g가 비공액 사전 분포인 로그정규분포인 경우 난수의 발생은 용이하지 않다. 이러한 경우 다른 방법으로 잠재적 변수를 활용하여 깁스표본에 적용이 가능한 방법이 Damien, Wakefield and Walker(1999)에 의하여 제시되었다. 다음절에서는 필요한 난수를 발생할 수 있 는 조건부 분포를 잠재적(latent) 변수를 도입하여 유도하였다.

4. 비공액 사전분포의 사용에서의 잠재적 변수 사용

앞장에서 보았듯이 λ_i의 조건부 분포함수로부터의 난수발생은 쉽지 않다. 이러한 경우 분포함수가 로그 오목함수라면 기각표본(rejection sampling) 방법을 사용할 수 있다. 그러나 이러한 방법조차 실제 적용은 쉽지 않다.

정리 4.1 :(Damien, Wakefield and Walker(1999)).
만약 다음의 함수 f로부터 난수를 얻으려면

$$f(x) \propto \pi(x)\prod_{i=1}^{N} l_i(x), \tag{4.1}$$

π는 알려진 분포함수이고 l_i는 양역함수(non-negative invertible function)이다.

그러면 함수 f로부터의 난수발생은 일양분포함수(uniform density)와 π의 절단 함수의 형태로부터 구할 수 있다.

위의 정리 4.1을 이용하여 조건부 분포를 유도할 수 있다.

모수 μ와 σ^2이 주어졌을 때 λ_1, λ_2, $\cdots$, λ_n들의 독립적인 사후분포함수는 다음밀도함수의 곱에 비례한다.

$$\prod_{i=1}^{n} f(y_i|\lambda_i)g(\lambda_i|\mu,\sigma^2) \propto \tag{4.2}$$

$$\prod_{i=1}^{n} \frac{e^{-exp(\lambda_i)}exp(\lambda_i)^{y_i}}{y_i!}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{(\lambda_i-\mu)^2}{2\sigma^2}\right\} \tag{4.3}$$

$$\propto \prod_{i=1}^{n} exp[-exp(\lambda_i)]exp\left\{-\frac{(\lambda_i-(\mu+\sigma^2 y_i))^2}{2\sigma^2}\right\}. \tag{4.4}$$

위의 식에서 잠재적 변수 $U = (u_1, u_2, \cdots, u_n)$를 도입하여 함수적 변환을 이용하면 λ_i와 U의 결합밀도함수는 다음과 같이 변환된다.

$$p(\lambda_1, \lambda_2, \cdots, \lambda_n, U|y, \mu, \sigma^2) \propto \tag{4.5}$$

$$\prod_{i=1}^{n} exp(-u_i) I\{u_i > exp(\lambda_i)\} exp\left\{-\frac{(\lambda_i - (\mu + \sigma^2 y_i))^2}{2\sigma^2}\right\}. \tag{4.6}$$

위의 λ_i와 U의 조건부 결합밀도함수에서 정리 4.1을 이용하여 각각의 조건부 확률밀도함수를 구하면 다음과 같다.

$$p(u_i|y, \mu, \lambda_{i;i=1,2,\cdots,n}, \sigma^2, u_{j;j\neq i}) \propto exp(-u_i) I\{u_i > exp(\lambda_i)\}, \tag{4.7}$$

$$p(\lambda_i|y, \mu, \sigma^2, \lambda_{j;j\neq i}, U) \sim N(\mu + \sigma^2 y_i, \sigma^2) I(\lambda_i < \log u_i). \tag{4.8}$$

앞에서 표현한 식(4.7)과 식(4.8)의 조건부 분포에서의 난수 발생이 용이하므로 MCMC 방법을 적용하여 (λ_i, U, μ, σ^2)의 결합 확률밀도 함수로부터 λ_i의 주변확률분포를 계산할 수 있다.

5. 모의실험과 결론

이 장에서는 포아송 분포를 갖는 자료에서 모수의 사전분포를 비공액 사전분포로 가정했을 때 추정 값을 앞 절에서 언급한 잠재적 변수를 사용하여 추정이 가능하다는 것을 모의 실험으로 보이겠다. 먼저, 자료에 대하여 언급하면 표준정규분포로부터 모수를 난수발생을 시키고 그 값을 이용하여 포아송 분포의 자료를 얻은 결과 다음의 표 5.1 과 같다.

모의실험 자료를 이용하여 모형을 정의하면 다음과 같다. 포아송-로그정규분포는 계층적 모형이므로 먼저 각각의 독립적 관찰치 y_i(i=1,2,...,10)는 포아송분포를 따른다고 가정하고 기대값은 e^{λ_i}를 갖는다고 가정하자. 여기서 모수 λ_i $(-\infty < \lambda_i < \infty)$에 관해서는 비공액 사전분포인 정규분포로 가정하자. 비공액 사전분포의 사용은 수학적 계산이나 조건부 분포함수의 복잡성을 유발시키지만 앞 절에서 언급한 잠재적 변수의 사용은 분포함수에서의 추정 계산을 쉽도록 도와준다. 또한, 모의 실험의 예를 보여주기 위하여 표준정규분포를 이용하였다. 각각의 모수 λ_1, λ_2, ... , λ_{10}들의 독립적인 사후분포함수는 식(4.2)를 이용하면 다음과 같다.

$$\prod_{i=1}^{10} \frac{e^{-exp(\lambda_i)} exp(\lambda_i)^{y_i}}{y_i!} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{{\lambda_i}^2}{2}\right\}. \tag{5.1}$$

표 5.1: 모의실험자료와 기대관찰치

순서	관찰치	모의실험 기대관찰치	모형에 의한 기대관찰치
1	0	1.09	0.53
2	0	0.77	0.49
3	2	2.61	1.72
4	0	0.20	0.50
5	0	0.38	0.52
6	2	1.56	1.87
7	4	3.18	3.57
8	0	0.41	0.49
9	0	0.34	0.51
10	3	1.11	3.73

식(5.1)에서 잠재적 변수 $U = (u_1, u_2, \cdots, u_{10})$를 도입하여 함수적 변환을 이용하고 식(4.7)과 식(4.8)과 같이 나타내면 각각의 조건부 확률밀도함수는 다음과 같다.

$$p(u_i|y, \lambda_{i;i=1,2,\cdots,n}, u_{j;j\neq i}) \propto exp(-u_i) I\{u_i > exp(\lambda_i)\}, \tag{5.2}$$

$$p(\lambda_i|y, \lambda_{j;j\neq i}, U) \sim N(y_i, 1) I(\lambda_i < \log u_i). \tag{5.3}$$

식(5.2)와 식(5.3)에서의 절단된 조건부 분포에서의 난수 발생이 용이하므로 MCMC방법을 적용하여 (λ_1, λ_2, $\cdots$, λ_{10}, U)의 결합확률밀도 함수로부터 각각 λ_i의 주변확률분포를 계산할 수 있다. 특히, 관찰치의 기대값이 초월함수에 비례하는 경우의 기대 관찰 추정값은 표 5.1에 있다. 위자료의 모의실험은 본 논문에서 제안한 모형이 성공적으로 수행되었는가에 관심을 가지고 결과를 계산 하였다. 하지만 잠재적 변수의 사용방법과 기각표본 사용방법의 효율성에 관해서는 비교는 하지 않았다. 결론적으로 본 논문의 초점은 포아송 모형에서의 비공액 사전분포의 사용에 관하여 제안하였다. 일반적으로 사전분포로써 공액 사전분포를 많이 사용하여 왔다. 공액 사전분포의 장점은 사후분포의 수학적 계산이 편리하고 공액 사전분포함수 형태의 집합이 상당히 크기 때문이다. 공액사전 분포를 사용하면 사후분포 역시 사전분포의 형태와 같거나 잘 알려진 분포함수의 형태를 나타낸다. 특히 최근에 사후분포의 계산을 하는 데 활용 성이 높은 MCMC 방법을 적용하려면 조건부 분포에서의 난수 발생이 용이하여야 한다. 그러기 위하여 공액 사전분포의 사용은 일반적인 경우로 인정되어 왔다. 그러나 자료의 형태에 따라 모형에서 모수의 조건에 충족하기 위하여 비공액 사전분포의 사용이 불가피한 경우가 발생한다. 본 논문에서는 포아송 모형에서 비공

액 사전분포인 로그정규분포의 사용도 MCMC 방법을 적용하는데 어렵지 않도록 하는 잠재적 변수 사용방법을 제시하였다. 그러나 기각표본 방법이나 또 다른 난수발생 방법이 더 효율적이라면 본 논문에서 제시한 잠재적 변수사용 방법을 활용하지 않아도 된다. 참고적으로 제시한 잠재적 변수 사용방법은 정리 4.1에서 보았듯이 어떠한 함수의 형태에서도 활용이 가능하다는 장점이 있다. 앞으로 필요하다고 생각하는 연구과제는 일반모형에 대하여 잠재적 변수사용, 기각표본, 그 외의 난수발생방법 등을 비교하여 계산의 효율성에 관하여 연구하고자 한다. 또한 이 논문에 대한 조언을 아끼지 않으신 익명의 편집위원에게도 감사를 드리고 싶다.

참고문헌

[1] Albert, J. and Chib, S. (1997). Bayesian Tests and Model Diagnostics in Conditionally Independent Hierarchical Models, *Journal of the American Statistical Association*, 92, 916-925.

[2] Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables, *Journal of the Royal Statistical Society*, Ser. B, 61, 331-344.

[3] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398-409.

[4] Geman, S. and Geman, D. (1984). Stochastic Relaxation Gibbs Distribution and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[5] Gelfand, A.E., Hill, S.E., Racin-poon, A. and Smith, A.F.G. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling, *Journal of the American Statistical Association*, 85, 972-985.

[6] Manton, K.G., Woodbury, M.A. and Stallard, E. (1981). A Variance Components Approach to Categorical Data Models With Heterogeneous Mortality Rates in North Carolina Counties, *Biometrics*, 37, 259-269.

[7] Tsutakawa, R.K. (1988). Mixed Model for Analyzing Geographic Variability in Mortality Rates, *Journal of the American Statistical Association*, 83, 37-42.

[1999년 11월 접수, 2000년 3월 채택]

A Study on Poisson-lognormal Model

Yong-Chul Kim[1)]

ABSTRACT

Conjugate prior density families were motivated by considerations of tractability in implementing the Bayesian paradigm. But we consider problem that the conjugate prior $p(\theta)$ cannot be used in restriction of the parameter θ. This article considers the non-conjugate prior problem of hierarchical Poisson model. We demonstrate the use of latent variables for sampling non-standard densities which arise in the context of the Bayesian analysis of non-conjugate by using a Gibbs sampler.

Keywords: Poisson-lognormal model; Monte Carlo method; Latent variable.

1) Assistant Professor, School of Computer and Information, Yongin University.
E-mail: yckim@eve.yongin.ac.kr