응용통계연구 (2011)
24(4), 677–693
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.4.677

영 과잉 포아송 모형에 대한 베이지안 방법 연구

이지호[1] · 최태련[2] · 우윤성[3]

[1]고려대학교 통계학과, [2]고려대학교 통계학과, [3]고려대학교 통계학과

(2011년 2월 접수, 2011년 6월 채택)

요 약

본 논문에서는 영 과잉 계수형 자료 분석을 위한 모형중의 하나인 영 과잉 포아송 모형의 베이지안 접근 방법에 대해서 연구한다. 구체적으로는 베이지안 영 과잉 포아송 모형의 적합을 위한 사후 표본을 추출하는데 있어서, 깁스 표집기(Gibbs sampler)를 이용하는 마르코프 연쇄 몬테칼로(MCMC) 방법과 역 베이즈공식(IBF)에 의한 표본추출 방법 두 가지를 고려한다. 이러한 두 가지 사후 표본 추출방법을 비교 설명하고, IBF를 통한 사후표본을 깁스 표집기 사후표본의 수렴성 여부를 확인하는 방식에 대해서도 소개한다. 이를 바탕으로 베이지안 영 과잉 포아송 모형을 Trajan이라는 사과 품종의 발아자료(Trajan data, Marin 등, 1993)에 적용하고 모수에 대한 사후추론을 실시하고 기존의 결과와 비교한다. 또한 주어진 자료에 대하여 영 과잉 포아송 모형이 적합한지에 대한 여부를 여러 가지 모형선택 기준을 통해서 살펴보고, 아울러 기존의 자료 분석 결과 (Rodrigues, 2003)를 보완하기 위하여 계층적 베이지안 모형과 같은 대안에 대해서도 논의해본다.

주요용어: 깁스 표집기, 역 베이즈 공식, 베이지안 카이제곱 적합도, DIC, 계층적 베이지안 모형.

1. 서론

셀 수 있는 값들을 갖는 계수형 자료(count data)는 보건 행정학, 의학, 사회학, 체육학, 공학 등 실생활과 관련된 여러 분야에서 관측될 수 있다. 이러한 계수형 자료 중에서 포아송 분포와 같은 일반적인 계수형 자료 분포에서 발생하는 영(zero)의 개수보다 더 많은 영 자료(zero data)가 발생하는 경우가 종종 있는데 이러한 자료를 분석하는 모형을 영 과잉 모형(Zero Inflated model)이라 한다. 이러한 영 과잉 모형은 실제 응용분야에서 계수형 자료를 분석하는데 있어서 다양하게 사용되어 왔다. 특히 포아송 분포 가정 하에서의 영 자료보다 더 많은 영 자료가 관측되는 모형을 영 과잉 포아송 분포(Zero Inflated Poisson distribution; ZIP)라고 한다 (Johnson 등, 2005). 영 과잉 포아송 분포는 사망률 자료와 같은 의학 통계분야에서 활용되어 왔으며, 예를 들어, 희귀질병으로 인한 사망 자료(mortality data due to rare diseases)에 대한 분석에 있어서는 ZIP 모형이 매우 유용하게 사용된다 (Ugarte와 Militino, 2004; Gómez-Rubio와 López-Quílez, 2010).

영 과잉 포아송 분포자료를 일반적인 포아송 모형으로 적합했을 때에는 영 과잉 부분에 대한 편향추정(biased estimation)이 발생하게 되고 이러한 문제를 해결하기 위해서 ZIP 모형을 이용하는 여러 가지 방법론이 제안되었고 실제 문제에 응용되었다. 예를 들어 Yip (1988)는 나뭇잎에 붙어있는 곤충의 개수를 분석하기 위하여 영 과잉 포아송(ZIP) 모형을 적용하였고 Lambert (1992)에 의해 ZIP 분포

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업이며 (No. 2010-0010422) 제 1저자 이지호의 석사학위논문, 이지호 (2011)의 계속 연구로 작성됨.
[2]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 부교수. E-mail: trchoi@korea.ac.kr

를 이용한 회귀분석 방법도 소개되었으며 Heilbron 등 (1989)은 영 과잉 포아송 분포와 더불어 이를 확장한 영 과잉 음이항 분포(Zero Inflated Negative Binomial distribution; ZINB)을 인간행동 연구에 적용하였다. 아울러 이러한 빈도론적인 방법론은 영 과잉 이항모형(Zero Inflated Binomial model), 영 과잉 음이항 모형(Zero Inflated Negative Binomial model), 영 과잉 일반화 포아송 모형(Zero Inflated Generalized Poisson model) 등으로 확대되었고 공변량을 포함시키는 다양한 영 과잉 회귀 모형(Zero Inflated Regression model)이 제안되었다. 베이지안적 관점에서는 Rodrigues (2003), Ghosh 등 (2006) and Angers와 Biswas (2003) 등에서 MCMC(Markov Chain Monte Carlo) 또는 Importance Sampling을 통한 Monte Carlo 방법론 등을 이용해 영 과잉 포아송(ZIP) 모형의 사후분포로부터 표본을 추출하고 모형을 적합하는 방법을 제시하였고 여러 가지 실제자료에 적용하였다. 이러한 베이지안 추론을 위해서 사용되는 MCMC와 같은 반복적 표본 추출 방법론(iterative sampling)은 표본 추출이 잘 이루어 졌는지에 대한 확인이 필요하며 반복추출 횟수를 수렴하기까지 충분히 크게 해야 하는 번거로움이 발생할 수 있다. 이와는 달리 역 베이즈 공식 표집기(Inverse Bayes Formula(IBF) sampler)를 통한 사후표본 추출방법 (Tan 등, 2003)은 정확 표집방법(exact sampling)으로서 조건부 표본 추출 방법이기 때문에 MCMC 방법과 같은 반복적 표본 추출 방법에서 발생하는 수렴 확인(convergence checking) 문제와 느린 수렴(slow convergence) 문제 혹은 많은 반복 횟수 문제를 피할 수 있다는 장점이 있다. 또한 IBF 표집을 통한 사후표본들은 MCMC 방법을 통한 사후표본들의 수렴여부를 확인하는데 활용될 수 있으며 (Tan 등, 2010), 선형혼합효과모형(linear mixed effects model)이나 결측자료 문제와 같은 다양한 응용문제에서도 정확 표집기로서 유용하게 활용되어 왔다 (Tian 등, 2007). 본 논문에서는 이러한 영 과잉 포아송 모형의 베이지안 분석을 위하여, MCMC를 통한 반복적 추출방법과 역 베이즈 공식 표집기에 의한 비 반복적 추출방법 두 가지를 고려하고 실제 응용문제에서의 자료에 대해 분석하고 그 결과를 논의한다. 구체적으로, 2절에서는 베이지안 영 과잉 포아송 모형 적합을 위한 사후 표본을 추출하는데 있어서, 깁스 표집기(Gibbs sampler)를 이용하는 MCMC 방법과 역 베이즈 공식 표집기(IBF sampler)를 이용하는 방법을 설명한다. 또한 깁스 표집기로부터 추출된 사후표본의 수렴성 여부를 확인하는 여러 가지 기준을 언급하고, 아울러 역 베이즈 공식 표집기를 통한 사후표본과 쿨백-라이블러 발산(Kullback-Leibler divergence)기준을 이용하여 깁스 표집기 사후표본의 수렴성 여부를 확인하는 방식에 대해서도 소개한다. 3절에서는 이를 바탕으로 베이지안 영 과잉 포아송 모형을 Trajan이라는 사과 품종의 발아에 관한 실제 자료(Trajan data, Marin 등, 1993)에 적용하고 모수에 대한 사후추론을 실시한다. 추가적으로 Trajan data에 대하여 영 과잉 포아송 모형이 적합한지에 대한 여부를 카이제곱 적합도, 사후확률검정 및 DIC(Deviance Information Criterion)을 사용하여 확인해보고 기존의 자료 분석 결과 (Rodrigues, 2003)에 대하여 비교 분석하고 보완하도록 한다. 아울러 기존의 자료 분석 결과 (Rodrigues, 2003)를 보완하기 위하여 계층적 베이지안 모형과 같은 대안에 대해서도 논의해본다. 끝으로 4절에서는 결론을 통해 본 논문을 정리하고 추가 연구의 방향에 대해서 논의한다.

2. 영 과잉 포아송 모형과 베이지안 적합 방법

2.1. 영 과잉 포아송 모형

이산형 확률 변수 Y_D가 모수 λ를 포함하는 $f(y;\lambda)$라는 확률 질량 함수를 가질 때, 모수 ϕ를 추가하여 다음과 같은 확률 질량 함수를 갖는 영 과잉 분포(zero inflated distribution; ZID), $Y \sim \text{ZID}(\phi, \lambda)$를 정의한다.

$$f(y;\phi,\lambda) = \phi_{I(y=0)} + (1-\phi)f(y;\lambda), \quad 0 \leq \phi \leq 1. \tag{2.1}$$

이 경우 영 과잉 확률변수 Y가 0일 확률은 $f(0;\phi,\lambda) = \phi + (1-\phi)f(0;\lambda)$이 되며 원래의 확률 변수

Y_D가 0일 확률 $f(0;\lambda)$보다 높음($\Pr(Y=0) \geq \Pr(Y_D=0)$)을 알 수 있고, 영 과잉 확률 변수 Y가 0보다 클 확률은 원래의 확률 변수 Y_D가 0보다 클 확률보다 작음($\Pr(Y>0) \leq \Pr(Y_D>0)$)을 쉽게 확인할 수 있다. 또한 영 과잉 분포는 영의 값만을 갖는 점 확률 분포와 일반적인 이산 확률 분포와의 혼합 분포(mixture distribution)라고 할 수 있다. 이러한 영 과잉 분포 중에서 가장 간단한 형태 중의 하나는 본 논문에서 고려하는 영 과잉 포아송 분포(Zero Inflated Poisson; ZIP)이며 식 (2.1)에서의 $f(y;\lambda)$를 포아송 확률 질량 함수 $e^{-\lambda}\lambda^y/y!$로 대체하여 다음과 같은 영 과잉 포아송 분포의 확률 질량 함수를 얻는다.

$$f(y;\phi,\lambda) = \phi I_{(y=0)} + (1-\phi)\frac{e^{-\lambda}\lambda^y}{y!}, \quad y=0,1,2,\ldots, 0 \leq \ \phi \leq 1, \tag{2.2}$$

앞서 언급했듯이 식 (2.2)의 영 과잉 포아송 분포는 영의 값만을 갖는 점 확률 분포와 포아송 분포의 혼합 분포로 간주 할 수 있고, 이 경우 혼합 비율(mixing proportion)은 모수 ϕ에 의해서 결정된다. 영 과잉 포아송 분포는 $\phi = 0$이면 일반적인 포아송 분포와 같은 분포가 되고 ϕ가 커질수록 더 많은 영 자료를 포함되게 된다. 따라서 일반적인 포아송 분포가 하나의 모수 λ만 갖는 것에 비해 영 과잉 포아송 분포는 두 개의 모수 ϕ와 λ를 갖는다. 따라서, 포아송 분포를 가정할 수 있는 이산형 자료를 분석하는데 있어서, 실제로 관측된 영 자료가 주어진 포아송 분포로부터 예상되는 영 자료보다 더 많은 경우에는, 일반적인 포아송 모형의 대안으로서 영 과잉 포아송 모형을 사용한다. 이러한 영 과잉 포아송 모형을 적합하기 위하여 먼저 모수 ϕ, λ의 가능도 함수(likelihood)를 알아보도록 한다 (Tan 등, 2010). $y_{obs} = (y_1, \ldots, y_n)$를 n개의 관측 값이라고 하면, 가능도 함수는 다음과 같다.

$$L(\phi,\lambda|y_{obs}) = [\phi + (1-\phi)e^{-\lambda}]^m \times (1-\phi)^{n-m}\Pi_{y_i \notin O}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}. \tag{2.3}$$

식 (2.3)에서 $O = \{y_i : y_i = 0, i = 1,2,\ldots,n\}$는 관측 값 y_i들 중에서 영 자료들의 집합이고 m은 집합 O의 원소의 개수를 의미한다. 집합 O의 원소, 즉 관측 값들 중에서 영의 값은 두 가지로 분류가 될 수 있는데, 하나는 영의 값만 갖는 분포로부터 얻은 경우와 포아송 분포에서 얻은 영의 경우 두 가지이다. 첫 번째 경우와 같은 상황에서, 영의 값만 갖는 분포로 관측 값에 작용을 한 Z라는 임의의 잠재 변수(latent variable)를 정의하면, m개의 영 관측 값 중에서 Z개의 영 관측값은 특정한 분포로부터 ϕ의 비율로 나오고 $m-Z$개의 영 관측 값은 포아송 분포로부터 $(1-\phi)e^{-\lambda}$의 비율로 나오는 이항분포를 따르게 되고, 성공확률은 $\phi/(\phi+(1-\phi)e^{-\lambda})$가 됨을 알 수 있다. 구체적으로는 ϕ, λ와 Y_{obs}가 주어졌을 때, Z에 대하여 다음과 같은 조건부 예측 분포(conditional predictive distribution)를 정의할 수 있다.

$$Z|y_{obs},\phi,\lambda \sim \text{Binomial}\left(m, \frac{\phi}{\phi+(1-\phi)e^{-\lambda}}\right). \tag{2.4}$$

식 (2.4)와 같이 잠재 변수를 도입하는 방식은 혼합분포(mixture distribution)의 적합을 위한 깁스 표집기나 EM 알고리즘 등에서 자료확대(data augmentation)을 통한 효율적인 계산을 위해서 사용된다 (예: Albert와 Chib (1993) 또는 Diebolt와 Robert (1994) 등). 이제 새롭게 도입된 잠재 변수 Z와 기존의 관측 값 Y_{obs}를 포함한 (ϕ,λ)에 대한 완전 자료, $D = (y_{obs}, z)$를 바탕으로 하는 가능도(complete-data likelihood) 함수를 정의하면 다음과 같다.

$$\begin{aligned} L(\phi,\lambda|D) &\propto \phi^z\left[(1-\phi)e^{-\lambda}\right]^{m-z} \times (1-\phi)^{n-m}\Pi_{y_i \notin O}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} \\ &\propto \phi^z(1-\phi)^{n-z}e^{-(n-z)\lambda}\lambda^{\sum_{y_i \notin O} y_i}. \end{aligned} \tag{2.5}$$

식 (2.5)에서 주어진 가능도 함수를 바탕으로 베이지안 적합을 하기 위해서는 적절한 사전분포(prior distribution)를 고려해야하며 이를 위하여 공액사전분포(conjugate prior distribution)를 사용하도록

한다. 가능도 함수의 형태를 살펴보면, 0과 1사이의 값을 갖는 모수 ϕ에 대해서는 이항분포의 공액 사전 분포로서의 Beta(a,b)를 사전 분포로 사용하는 것이 합당하고, 포아송 분포의 모수인 λ에 대해서는 Gamma(c,d)를 쓰는 것이 합당하다고 여겨지며 이 두 사전분포는 서로 독립이라고 가정한다.

$$\phi \sim \text{Beta}(a,b), \qquad \lambda \sim \text{Gamma}(c,d). \tag{2.6}$$

따라서 식 (2.5)의 가능도 함수와 식 (2.6)의 사전분포를 결합하여 모수 (ϕ,λ)에 대하여 다음과 같은 결합 확대 사후분포(joint augmented posterior distribution)를 얻게 된다.

$$\begin{aligned} p(\phi,\lambda|D) &\propto L(\phi,\lambda|D)\times p(\phi)p(\lambda) \\ &\propto \phi^{z}(1-\phi)^{n-z}e^{-(n-z)\lambda}\lambda^{\sum_{y_i\notin O} y_i}\times \phi^{a-1}(1-\phi)^{b-1}\lambda^{c-1}e^{-d\lambda}. \end{aligned} \tag{2.7}$$

또한 식 (2.7)의 결합 확대사후분포는 λ와 ϕ 두 확대사후분포의 곱으로 표현될 수 있으며 각 사후분포는 식 (2.8)과 같이 계산됨을 알 수 있다.

$$\begin{aligned} \phi|D &\sim \text{Beta}(z+a, n-z+b), \\ \lambda|D &\sim \text{Gamma}\left(\sum_{y_i\notin O} y_i + c, n-z+d\right). \end{aligned} \tag{2.8}$$

결론적으로 영 과잉 포아송 분포(ZIP)에 대한 가능도 함수를 바탕으로 적절한 사전분포 하에서 식 (2.8)의 사후분포를 유도하였다. 다음 절에서는 이를 바탕으로 사후표본을 추출하여 베이지안 추론을 시행하는데 있어서, 우리가 고려해 볼 수 있는 두 가지 수치적 방법, 깁스 표집기 알고리즘과 역 베이즈 공식 표집기 알고리즘에 대해서 구체적으로 설명해보도록 한다.

2.2. 깁스 표집기(Gibbs sampler)를 통한 사후 표본 추출

2.1절의 식 (2.8)에서 설명된 사후분포는 잠재변수 Z를 포함하는 완전자료$D=(y_{obs},z)$를 바탕으로 유도된 것이므로 이를 통한 사후 표본추출에서는 잠재변수 Z에 대한 표본추출이 필요하다. 깁스 표집을 통한 사후표본 추출에서는 잠재변수 Z에 대한 표본추출을 위해서 식 (2.4)에서 설명된 이항분포를 따르는 완전 조건부(full conditional)분포를 이용한다. 이를 바탕으로 완전자료를 구성하고 식 (2.8)의 사후분포로부터 ϕ와 λ에 관한 사후 표본을 추출하도록 한다. 구체적으로 j번째 $(j=1,2,\ldots)$ 사후표본을 추출하기 위해서 다음과 같은 절차를 거친다.

- 단계 1: $(j-1)$번째 사후표본 추출 값 $(\phi^{(j-1)},\lambda^{(j-1)})$으로부터 얻어진 식 (2.9)와 같은 이항분포에서 표본 $Z^{(j)}$를 추출해낸다.

$$Z^{j}|y_{obs},\phi^{(j-1)},\lambda^{(j-1)} \sim \text{Binomial}\left(m, \frac{\phi^{(j-1)}}{\left[\phi^{(j-1)}+(1-\phi^{(j-1)})e^{-\lambda^{(j-1)}}\right]}\right) \tag{2.9}$$

- 단계 2: 단계 1에서 얻은 표본 $Z^{(j)}$를 식 (2.8)의 사후분포에 대입하여 얻은 각각의 사후분포로부터 모수에 대한 표본 $\phi^{(j)}$와 $\lambda^{(j)}$를 추출해낸다.

$$\begin{aligned} \phi^{(j)} &\sim \text{Beta}\left(Z^{(j)}+a, n-Z^{(j)}+b\right), \\ \lambda^{(j)} &\sim \text{Gamma}\left(\sum_{y_i\notin O} y_i + c, n-Z^{(j)}+d\right) \end{aligned} \tag{2.10}$$

이렇게 얻은 각 단계에서의 표본들을 기록해두고 정해놓은 반복 횟수까지 계속해서 반복 추출해낸다. 이 경우, 표본 추출 후 확인해봐야 할 사항은 사후 표본의 수렴성(convergence) 여부이다. 깁스 표집기를 이용한 - 일반적으로는 MCMC 방법을 통한 - 사후표본 추출에서는, 표본을 추출하기 위한 각각의 마르코프 연쇄(Markov chain)의 정상 분포(stationary distribution)가 우리가 원하는 사후분포인지, 즉 마르코프 연쇄가 사후분포로 수렴하는지 확인해야한다. 만약 수렴하지 않았다면 수렴할 때까지 반복횟수를 늘여야 하고 적당한 소각표본(burn-in sample)을 제외해야 한다. 처음의 반복 횟수보다 더 많은 반복 횟수를 필요로 하게 되며 수렴이 느리게 되는 문제(slow convergence problem)를 야기하기도 한다. 이러한 수렴여부 진단(convergence diagnostics)을 위해서는 경험적으로 트레이스 그림(trace plot)이나 두 개 이상의 병렬 연쇄를 사용하여 확인할 수 있다. 또한 수치적으로는 Gelman-Rubin 통계량 (Rubin과 Gelman, 1992; Brooks와 Gelman, 1998) 등을 이용하여 수렴여부를 진단할 수 있다.

2.3. 역 베이즈 공식(Inverse Bayes Formula; IBF)을 통한 사후 표본 추출

역 베이즈 공식 표집기(IBF sampler)는 2.2절의 깁스 표집기와는 달리 역 베이즈 공식을 이용한 정확표집(exact sampling)을 통해 표본을 추출한다 (Tian 등, 2007; Tan 등, 2010). 베이즈 공식의 역방향을 고려하는 역 베이즈 공식 (Tan 등, 2003, 2010)은 다음과 같이 두 조건부 확률밀도함수의 비를 통해 주변확률 밀도함수를 얻게 되는 공식을 의미한다.

$$f_Y(y) = \frac{f_{Y|X}(y|x)}{f_{X|Y}(x|y)} \cdot f_X(x), \qquad f_X(x) = \left[\int \frac{f_{Y|X}(y|x)}{f_{X|Y}(x|y)} dy\right]^{-1} \tag{2.11}$$

$$\Longrightarrow f_X(x) = \left[\int \frac{f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)} dx\right]^{-1} \cdot \frac{f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)} \propto \frac{f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)}. \tag{2.12}$$

식 (2.11)로부터 X와 Y를 바꾸어 놓고 임의로 주어진 값 y_0를 대입하면 식 (2.12)를 얻게 된다 (Tan 등, 2010). 이러한 역 베이즈 공식을 바탕으로 제안된 역 베이즈 공식 표집기는 EM(Expectation Maximization)/DA(Data Augmentation) 알고리즘과 유사한 형태로 사용되며, 베이지안 추론에서는 확대사후분포(augmented posterior distribution)나 조건부 예측분포(conditional predictive distribution)를 통한 사후표본 추출에서 주로 사용된다 (Tan 등, 2003). 역 베이즈 공식 표집기를 이용한 모형적합 및 사후표본 추출은 결측(missing) 값을 포함하는 이산형 모형에서도 사용되었으며 (Tian 등, 2007), 본 논문에서 고려하는 영 과잉 포아송 모형에서도 쉽게 적용이 되며, 식 (2.4)에서 도입된 잠재변수의 조건부 예측분포와 식 (2.7)과 (2.8)의 확대된 사후분포를 이용하여 역 베이즈 공식 표집기를 구성한다. 보다 구체적으로 역 베이즈 공식 표집기의 절차를 설명하면 다음과 같다. 2.2절에서 설명된 깁스 표집기를 이용하는 절차에서와 마찬가지로 잠재변수 Z에 대한 표본추출이 먼저 필요하며 역 베이즈 공식 표집기에서는 깁스 표집기와는 달리 잠재 변수의 조건부 분포 $f_{Z|Y_{obs}}(z|Y_{obs})$으로부터 표본 $\{Z^{(l)}\}_{l=1}^{L}$을 생성한다. l번째 표본 $Z^{(l)}$과 관측값 Y_{obs}를 결합한 완전자료 $D^{(l)} = (Y_{obs}, Z_l)$ 이용한 $f(\phi, \lambda|D^{(l)})$ 분포에서부터 추출한 표본 $\{(\phi^{(l)}, \lambda^{(l)})\}_{l=1}^{L}$을 얻을 수 있다. 이 때 추출된 표본$\{(\phi^{(l)}, \lambda^{(l)})\}_{l=1}^{L}$은 사후분포 $f(\phi, \lambda|Y_{obs})$로부터 얻은 독립이고 동일한 분포로부터 얻은 표본이라고 할 수 있고 $f(\phi, \lambda|Y_{obs}) = \int f(\phi, \lambda|Y_{obs}, z) f_{Z|Y_{obs}}(z|Y_{obs}) dz$의 관계로 설명이 가능하다. 따라서 역 베이즈 공식 표집기의 관건은 잠재 변수 Z에 대한 표본을 구하는 일이다. 식 (2.11)과 (2.12)에서 X의 주변확률밀도 $f_X(x)$를 보다 일반적으로 조건부 확률밀도함수를 이용해서 나타낼 수 있으며, 구체적으로 Y_{obs}가 주어졌을 때의 Z에 대한 조건부 예측 확률밀도함수 $f_{Z|Y_{obs}}(z|Y_{obs})$로 대체하고, y_0를 임의의 주어진 모수값 (ϕ_0, λ_0)로 대체하면 아래의 식 (2.13)을 얻게된다 (Tan 등, 2003, 2010).

$$f_{Z|Y_{obs}}(z|Y_{obs}) = \left[\int \frac{f_{Z|Y_{obs},\phi,\lambda}(z|Y_{obs}, \phi_0, \lambda_0)}{f_{\phi,\lambda,Y_{obs}|Z}(\phi_0, \lambda_0, Y_{obs}|z)} dz\right]^{-1} \frac{f_{Z|Y_{obs},\phi,\lambda}(z|Y_{obs}, \phi_0, \lambda_0)}{f_{\phi,\lambda|Y_{obs},Z}(\phi_0, \lambda_0|Y_{obs}, Z)}$$

$$\propto \frac{f_{Z|Y_{obs},\phi,\lambda}(z|Y_{obs},\phi_0,\lambda_0)}{f_{\phi,\lambda|Y_{obs},Z}(\phi_0,\lambda_0|Y_{obs},z)}. \tag{2.13}$$

이 경우, 잠재 변수 Z가 이산형 확률변수이므로 $f_{Z|Y_{obs}}(z_k|Y_{obs}) = \Pr\{Z = z_k|Y_{obs}\} = p_k$로 대체되고, Z의 분포를 알기 위해서는 확률질량함수 $p_k = f(z_k|Y_{obs})$, $k = 1,\ldots,K$를 구하도록 한다. 구체적으로 식 (2.13)의 역 베이즈 공식을 활용하면 식 (2.14)의 q_k값들을 얻게된다,

$$q_k = q_k(\phi_0,\lambda_0) = \frac{\Pr\{Z = z_k|Y_{obs},\phi_0,\lambda_0\}}{f(\phi_0,\lambda_0|Y_{obs},z_k)}, \quad k = 1,\ldots,K. \tag{2.14}$$

식 (2.13)의 역베이즈 공식에서 주어진 모수 값 (ϕ_0,λ_0)은 일반적으로 임의로 결정되며, 이를 이용한 q_k의 값은 (ϕ_0,λ_0)에 의존하며, 식 (2.14)와 같이 $q_k(\phi_0,\lambda_0)$로 표현된다. 따라서, 역베이즈 공식은 이론적으로는 임의의 모수값에 대해 항상 성립하지만 실제적으로는 주어진 모수 값의 영향을 상당히 받는 것으로 알려져 있다. 특히 식 (2.14)에서 분모에 밀도함수가 들어가는데 이 밀도함수값이 0에 가까운 경우에 가중치(weight) q_k를 지나치게 크게 만들고 따라서 소수의 q_k가 매우 커져 q_k의 변동성(variation)을 크게 만들 수 있는 단점이 있다. 이 경우 $f_{Z|Y_{obs}}(z|Y_{obs})$에 대하여 $f_{Z|Y_{obs},\phi,\lambda}(z|Y_{obs},\phi_0,\lambda_0)$를 통한 근사값 중에서 최적인 값(best approximation)에 대응하는 ϕ_0와 λ_0를 선택하도록 한다 (Tan 등, 2003). 실제 계산을 위해서는 사후 밀도 함수 $f(\phi,\lambda|Y_{obs})$를 최대로 하는 값 $(\widehat{\phi}_{obs},\widehat{\lambda}_{obs})$, 즉 사후 분포 최빈값(posterior mode)을 선택하면 충분하다는 것이 알려져 있다 (Tan 등, 2003, Theorem 1).

이제, Z의 조건부 확률질량함수 또는 실제 확률에 대응하는 p_k를 다음과 같이 구한다. 식 (2.14)로부터 얻어진 q_k값들을 전체 대비 비율로 표준화하여, 즉 총합으로 나누어 계산하면 식 (2.15)의 p_k값들을 얻게 된다,

$$p_k = \frac{q_k(\phi_0,\lambda_0)}{\sum_{k=1}^{K} q_k(\phi_0,\lambda_0)}, \quad k = 1,\ldots,K. \tag{2.15}$$

식 (2.15)의 확률 값 p_k는 이산확률 변수 Z의 조건부 확률 질량함수 $f(Z|Y_{obs})$를 결정하며, 이를 바탕으로 표본 $\{Z^{(l)}\}_{l=1}^{L}$를 표집한다. 따라서 j번째 $(j = 1,2,\ldots)$ 사후표본을 추출하기 위한 과정들을 요약하면 다음과 같다.

- 단계 1: 임의로 주어진 (ϕ_0,λ_0)로부터 식 (2.4)의 이항확률을 계산하고 식 (2.7)의 사후밀도함수를 계산하고 두 값의 비율 q_k를 계산한다.

$$\Pr(Z = k|y_{obs},\phi_0,\lambda_0) = \binom{m}{k}\psi_0^k(1-\psi_0)^{m-k}, \quad k = 0,1,\ldots,m, \quad \psi_0 = \frac{\phi_0}{[\phi_0 + (1-\phi_0)e^{-\lambda_0}]}$$

$$f(\phi_0,\lambda_0|D) \propto \phi_0^k(1-\phi_0)^{n-k}e^{-(n-k)\lambda_0}\lambda_0^{\sum_{y_i\notin O} y_i} \times \phi_0^{a-1}\left(1-\phi_0^{b-1}\right)\lambda_0^{c-1}e^{-d\lambda_0}$$

$$q_k = q_k(\phi_0,\lambda_0) = \frac{\Pr\{Z = k|Y_{obs},\phi_0,\lambda_0\}}{f(\phi_0,\lambda_0|Y_{obs},k)}, \quad k = 0,1,\ldots,m.$$

- 단계 2: 단계 1에서 얻은 q_k를 표준화하여 Z의 조건부 확률질량함수를 얻는다.

$$f(Z = k|Y_{obs}) = p_k = \frac{q_k(\theta_0)}{\sum_{k=0}^{m} q_k(\theta_0)}, \quad k = 0,1,\ldots,m.$$

- 단계 3: 단계 2에서 얻은 p_k를 확률분포로 갖는 이산형 확률변수 Z로부터 L개의 표본 $\{Z^{(l)}\}_{l=1}^{L}$을 추출한다.

- 단계 4: 단계 3에서 얻은 표본 $Z^{(l)}$을 식 (2.8)의 사후분포에 대입하여 얻은 각각의 사후분포로부터 모수에 대한 표본 $(\phi^{(l)}, \lambda^{(l)})$을 추출한다.

$$\phi^{(l)} \sim \text{Beta}\left(Z^{(l)} + a, n - Z^{(l)} + b\right),$$
$$\lambda^{(l)} \sim \text{Gamma}\left(\sum_{y_i \notin O} y_i + c, n - Z^{(l)} + d\right).$$

이 경우, 역 베이즈 공식 표집기를 통해 추출된 사후표본 $\{(\phi^{(l)}, \lambda^{(l)})\}_{l=1}^{L}$은 비 반복적인 정확표집(noniterative exact sampling)이며 추출횟수 L을 증가시킴으로써, 실제 사후분포와 가깝게 된다. 따라서 반복적인 추출이 아니기 때문에 깁스 표집기를 통한 사후 표본과 같은 수렴성 여부를 조사할 필요가 없다는 장점이 있으며, 또한 깁스 표집기를 통한 사후 표본의 수렴성 조사에 있어서 하나의 기준(benchmark)으로 활용될 수 있다. 구체적으로는 깁스 표집기와 역 베이즈 공식 표집기로부터 얻은 사후표본들을 이용하여, 두 표본들 사이의 무질서도(entropy)가 얼마나 일치하는지를 쿨백 라이블러 발산기준(Kullback-Leibler divergence)을 통해 비교할 수 있다. 이 경우 두 사후표본의 무질서도의 비(ratio)에 로그를 취한 값이 쿨백 라이블러 발산으로 정의되므로, 무질서도가 비슷해질수록 그 비율은 1이 되고 쿨백 라이블러 발산 값은 영에 가까운 값을 갖게 된다는 사실을 바탕으로 두 사후표본을 비교할 수 있다 (Tan 등, 2010). 반면에, 역 베이즈 공식 표집기를 사용하는 경우에는, 깁스 표본기법에 비해 수렴성 문제가 없는 대신, 이전에 설명했던 바와 같이, q_k의 안정성(stability)을 확인하는 것이 필요하다. 이러한 문제를 해결하기 위해서, 경험적으로, 여러가지 주어진 모수값 (ϕ_0, λ_0)을 이용하는 표집기 중에서 가장 안정적인, 즉 변동성이 작은 값을 선택하거나, 사후분포의 최빈값을 주어진 모수값 (ϕ_0, λ_0)으로 사용하는 역 베이즈 공식 표집기를 사용하도록 한다. 이러한 사실을 바탕으로, 두 사후 표본에 대한 실제 자료를 통한 실증적인 비교는 다음 절에서 소개된다.

3. 실제 자료 적용: Trajan 자료에 대한 분석 및 고찰

3절에서는 앞 절에서 논의된 깁스 표집기와 역 베이즈 공식 표집기를 이용하여 베이지안 영 과잉 모형에서의 사후분포로부터 표본을 추출하고 이를 바탕으로 통계적 추론을 실시한다. 베이지안 영 과잉 모형을 통한 자료 적합을 위하여 Trajan이라는 사과 품종의 발아에 관한 실제 자료(Trajan data, Marin 등, 1993)에 적용하고 기존의 자료 분석 결과 (Rodrigues, 2003)와 비교 분석하고 보완하도록 한다. 또한 Trajan 자료가 영 과잉 포아송 모형이 적합한지의 여부를 베이지안 모형 선택 및 가설 검정의 관점에서 살펴보고, 역 베이즈 공식 표집기 사후 표본을 이용하여 깁스 표집기 사후 표본의 수렴여부와 역 베이즈 공식 표집기의 안정성에 대해서도 살펴본다.

3.1. Trajan 자료에 대한 선행 연구: Rodrigues (2003)

표 3.1에 제시된 Trajan 자료는 사과 품종 Trajan의 발아된 싹의 개수를 기록한 자료로서 Marin 등 (1993)의 연구 결과로부터 제공된다. 이러한 Trajan 자료를 분석하기 위하여 기존의 자료 분석 (Rodrigues, 2003)에서는 베이지안 포아송 영 과잉 모형을 적용하였다. 구체적으로, 깁스 표집기를 이용한 MCMC 방법을 통하여 포아송 영 과잉 모형의 사후분포를 통한 통계적 추론을 실시하였다. 표 3.1의 Trajan 자료(또는 Trajan 사과 발아 자료)는 미소대량증식(mircopropgation)을 통해 발아된 Trajan 품종 270 싹(shoot)에 대한 뿌리(root) 개수의 빈도를 제공한다. 이러한 실험을 위하여, 광주기(빛을 쏘여준 시간, photoperiod)와 BAP라는 식물생장호르몬(cytokinin)의 농도 두 가지 인자(factor 또는

표 **3.1.** Apple cultivar Trajan root data

BAP	Photoperiod									
	8					16				
	2.2	4.4	8.8	17.6	Total	2.2	4.4	8.8	17.2	Total
roots										
0	0	0	0	2	2	15	16	12	19	62
1	3	0	0	0	3	0	2	3	2	7
2	2	3	1	0	6	2	1	2	2	7
3	3	0	2	2	7	2	1	1	4	8
4	6	1	4	2	13	1	2	2	3	8
5	3	0	4	5	12	2	1	2	1	6
6	2	3	4	5	14	1	2	3	4	10
7	2	7	4	4	17	0	0	1	3	4
8	3	3	7	8	21	1	1	0	0	2
9	1	5	5	3	14	3	0	2	2	7
10	2	3	4	4	13	1	3	0	0	4
11	1	4	1	4	10	1	0	1	0	2
12	0	0	2	0	2	1	1	1	0	3
>12	13, 17	13	14, 14	14	6					

실험조건)를 조절하고 이에 따라서 발아된 각각의 싹에 대한 뿌리의 개수를 기록하였다. 광주기와 BAP 농도, 두 가지 인자에 대하여, 첫 번째로 광주기를 8시간과 16시간으로 구분하여 두 집단으로 나누었고, 두 번째로 각 집단에 대해 BAP의 농도를 4 단계(2.2, 4.4, 8.8, 17.2)로 나누어 발아된 싹에서의 뿌리의 개수를 계측하였다. 그 결과는 표 3.1에 기록된 자료와 같다.

자료 분석을 위해서 먼저 Rodrigues (2003)와 마찬가지로 BAP 농도별 4단계 실험조건은 고려하지 않고 광주기에 따른 두 그룹(8시간/16시간)의 자료를 분석하기로 한다. 표 3.1에서 BAP 농도에 따른 차이가 병합된 총계(Total)에 해당하는 뿌리의 개수가 관심 있는 계수형 자료(count data)가 된다. 특히 광주기(photoperiod)가 16시간인 경우, 뿌리의 개수가 0인 싹의 개수가 62개로서 영 과잉 자료임을 쉽게 확인할 수 있다. 따라서, 광주기가 16시간인 자료가 주 관심 대상이라고 하겠다. Rodrigues (2003)에서는 두 계수형 자료를 분석하기 위하여 포아송(P) 모형과 영 과잉 포아송(ZIP) 모델을 사용하여 적합한 결과를 제시하였고 광주기 16시간의 적합 결과는 표 3.2와 같다.

E_P와 E_{ZIP}는 각각 포아송(P) 모형과 영 과잉 포아송(ZIP)을 사용하여 적합한 빈도(Fitted frequency)를 나타내며 영 관측 값의 경우 포아송 모형에 비해서 영 과잉 포아송 모형이 잘 적합함을 알 수 있으며 그 외의 관측 값에 대해서는 포아송 모형이 과다추정의 경향을 영 과잉 포아송 모형이 과소 추정의 경향을 보임을 알 수 있다. 또한 Rodrigues (2003)에서는 포아송 모형과 영 과잉 포아송 모형(ZIP)간의 선택을 위한 기준으로 $\chi^2 = \sum_i (O_i - E_i)^2/E_i$값을 제공하였으며 표 3.2에 제시된 광주기 16시간 자료에 대한 두 적합 값, 2954와 49.85를 통해서 볼 때, 광주기(photoperiod) 16시간의 경우는 ZIP 모형이 선호됨을 알 수 있다.

3.2. 실제 자료에 대한 응용(Trajan Data) 및 분석 결과

본 논문에서의 Trajan 자료에 대한 분석에서는 Rodrigues (2003)에서 행해진 기존의 분석 결과를 비교하고 보완하도록 한다. 구체적으로, Trajan 자료를 분석하기 위한 베이지안 ZIP 모형을 적합하는데 있어서 앞 절에서 설명한 깁스 표집기와 역 베이즈 공식 표집기를 사용하여 사후표본을 추출하고 이를 바

표 **3.2.** Rodrigues (2003)의 적합 결과: 광주기 16기간

No. of roots	Obs	E_P	E_{ZIP}
0	62	7.43	61.68
1	7	21.27	1.74
2	7	30.43	4.63
3	8	29.02	8.23
4	8	20.76	11.00
5	6	11.88	11.80
6	10	5.66	10.57
7	4	2.31	8.13
8	2	0.82	5.49
9	7	0.26	3.31
10	4	0.07	1.79
11	2	0.01	0.89
12	3	0.00	0.40
		$\chi^2_P = 2954$	$\chi^2_{ZIP} = 49.85$

탕으로 다음과 같은 관점에서 자료 분석을 실시한다. 첫째, 깁스 표집기를 통한 사후표본의 추출에 있어서, 본 논문에서는 실제 베이지안 자료 분석에서 많이 사용되는 WinBUGS (Spiegelhalter 등, 2003) 프로그램을 이용한다. WinBUGS에서는 베이지안 사후추론에서 필요한 다양한 통계량을 제공함으로써 사용자에게 편의를 제공한다 (Ntzoufras, 2009). 예를 들어 WinBUGS에서는 사후표본의 수렴 여부 확인을 위하여 2절에서 언급되었던 트레이스 그림(trace plot)이나 Gelman-Rubin 통계량을 자동적으로 제공하기 때문에 이를 이용하여, 수렴여부 확인이 간략히 언급된 Rodrigues (2003)의 결과들을 보완한다. 둘째, 역 베이즈 공식 표집기를 통한 사후표본을 이용하여 깁스 표집기를 통한 사후표본의 결과와 비교한다. 또한 2절에서 설명된 것처럼 깁스 표집기를 통한 사후표본의 수렴성 여부를 역 베이즈 공식 표집기 사후표본을 이용하는 대안적인 절차를 바탕으로 Trajan 자료를 적용하고 실증적으로 확인하여 Rodrigues (2003)의 결과를 보완한다. 셋째, ZIP 모형선택을 위하여 Rodrigues (2003)에서 사용한 전통적 χ^2적합도 검정기준 뿐만 아니라 사후확률을 바탕으로 하는 베이지안 검정을 실시하고, 아울러 베이지안 모형 적합에서 실용적으로 자주 사용되며 동시에 WinBUGS에서 제공되는 DIC(Deviance Information Criterion, Spiegelhalter 등, 2002)를 이용하여 모형 선택 기준으로 활용하도록 한다. 마지막으로, Trajan 자료를 적합하는데 있어서 Rodrigues (2003)에서 고려하지 않았던 BAP 농도에 따른 차이를 고려한, 계층적 베이지안(hierarchical Bayesian) 영 과잉 모형에 대해서도 논의해본다. 이러한 관점을 바탕으로 베이지안 영 과잉 포아송 모형을 적합하기 위하여 식 (3.1)에 주어진 모형구조와 무정보적(noninformative) 사전분포를 이용한다.

$$\begin{aligned} Y_i &\sim \text{ZIP}(\phi, \lambda), \quad i = 1, \ldots, n, \\ \phi &\sim \text{Beta}(0.5, 0.5), \\ \lambda &\sim \text{Gamma}(0.001, 0.001). \end{aligned} \tag{3.1}$$

베이지안 ZIP 모형에서는, 일반적으로 0과 1사이의 값을 갖는 모수 ϕ에 대해서는 $\phi \sim \text{Beta}(a, b)$를 사용하고 포아송 모수 λ에 대해서는 $\lambda \sim \text{Gamma}(c, d)$의 형태를 사용한다 (Tan 등, 2003; Ghosh 등, 2006). 이 경우 각각의 사전분포는 Bernoulli(ϕ)와 Poisson(λ), 일반적으로는 Power series(PS) 분포의 공액사전분포로 이해할 수 있다 (Ghosh 등, 2006). 기존의 자료분석 결과인 Rodrigues 등 (2003)에서는 $a = b = c = 1/2$와 $d = 0$를 갖는 다소 불명확한 사전분포가 명시되어 있다. 본 논문에서는

표 **3.3.** (ϕ_0, λ_0)에 따른 q_k의 변동성 비교

(ϕ_0, λ_0)	$(0.5, 4.2)$	$(0.55, 4.4)$	$(0.4, 5)$	$(\widehat{\phi}_{obs}, \widehat{\lambda}_{obs}) = (0.4743, 5.4314)$	$(0.45, 5.6)$
$\mathrm{Var}(q)$	289275.9	4291.446	0.01044	0.000058	0.00011

표 **3.4.** 깁스 표집기 (WinBUGS 이용)와 역 베이즈 공식 표집기를 통한 사후 표본 요약

	사후표본	mean	s.d	2.5	25	median	75	97.5
λ	깁스표집기(WinBUGS)	5.441	0.279	4.906	5.253	5.440	5.625	6.003
	역베이즈공식표집기(R)	5.447	0.279	4.927	5.247	5.431	5.637	5.990
ϕ	깁스표집기(WinBUGS)	0.475	0.042	0.393	0.446	0.474	0.505	0.556
	역베이즈공식표집기(R)	0.476	0.044	0.391	0.445	0.475	0.504	0.567

무정보적 사전분포(분산이 매우 큰, high variance)를 고려하기 위해서, 식 (3.1)에서는 $a = b = 0.5$, $c = d = 0.001$을 사용하였으며, 이러한 무정보적 사전분포는 Ghosh (2006, Section 3)에서도 고려되었다.

표 3.4은 식 (3.1)의 베이지안 영 과잉 포아송 모형의 사후 분포에 대한 추론을 위하여 깁스 표집기와 역 베이즈 공식 표집기를 통해 추출한 1000개의 사후표본에 대한 요약 통계량 (평균, 표준편차, 0.025, 0.25, 0.5, 0.75, 0.975 분위수)을 나타낸다. 깁스 표집기를 통한 사후 표본추출을 위하여 WinBUGS 프로그램(부록의 코드 참조)을 이용하였고 사후표본추출에 있어서는 각각의 모수별로 5000번의 반복시행을 하였고 4000개의 소각샘플(burn-in sample)을 제거하였다. 깁스 표집기의 수렴여부를 확인하기 위하여 R-package 중의 하나인 CODA(Convergence Diagnosis and Output Analysis)에서 제공하는 다양한 진단 통계량 및 검정(Geweke diagnostic (Geweke, 1992), Gelman-Rubin diagnostic (Gelman과 Rubin, 1992), Raftery-Lewis diagnostic (Raftery와 Lewis, 1992))을 이용하여, 세 개의 병렬체인을 사용하는 WinBUGS를 통해 추출한 MCMC 결과물에 대한 수렴여부를 확인하였다. 또한 WinBUGS에서 자체적으로 제공하는 다중 트레이스 그림과 BGR diagnostic (Gelman과 Rubin, 1992; Brooks와 Gelman, 1998)을 이용하여 ϕ와 λ 두 모수에 대한 사후표본의 수렴여부를 확인하였다. 수렴 여부를 확인하기 위한 다양한 기준을 통해서 살펴 본 결과 깁스 표집기를 통한 사후표본은 각각의 사후분포로 수렴함을 알 수 있었다.

2절에서 설명되었던 것처럼, 역 베이즈 공식 표집기를 사용하는 경우, 주어진 모수 값 (ϕ_0, λ_0)를 결정하기 위해서 사후 분포 최빈값을 사용하는 것이 바람직하며, 본 자료분석에서는 식 (3.1)의 베이지안 모형으로 부터 결정되는 사후 분포 최빈값(posterior mode)을 사용하였다. 이를 위해서 R library 중의 하나인 LearnBayes (Albert, 2007)에서 제공하는 `laplace`함수를 사용하여 수치적으로 사후 분포 최빈값을 계산하였다. 이를 통하여, 사후 분포 최빈값은 $(\widehat{\phi}_{obs}, \widehat{\lambda}_{obs}) = (0.4743, 5.4314)$로 얻어졌고, 이 값을 (ϕ_0, λ_0)로 사용하였다. 또한 2절에서 설명되었던 것처럼, q_k의 값은 (ϕ_0, λ_0)에 의존하며, (ϕ_0, λ_0)의 값에 따라 q_k의 변동성을 크게 만들 수 있는 단점이 있다. 이를 보완하기 위하여, (ϕ_0, λ_0)값에 따른 q_k의 변동성을 $\mathrm{Var}(q)$로 측정하였고, 이 경우 사후 분포 최빈값 $(\widehat{\phi}_{obs}, \widehat{\lambda}_{obs})$을 (ϕ_0, λ_0)로 사용했을 때의 q_k의 변동성이 가장 작음을 확인할 수 있었다. 표 3.3은 이러한 수치적 비교를 위한 계산 값들 중, 몇 가지 $\mathrm{Var}(q)$값들을 그러한 예로서 제시한 결과이다. 표 3.4의 역 베이즈 공식 표집기를 통한 결과는 사후 분포 최빈값 $(\widehat{\phi}_{obs}, \widehat{\lambda}_{obs})$을 이용한 결과이다.

표 3.4의 역 베이즈 공식 표집기를 사용한 결과는 깁스 표집기를 통한 결과와 비교하기 위해서 1000개의 사후표본을 추출하였다. 병렬체인을 사용하고 소각과정을 거쳐야하며 또한 수렴여부를 확인 해야만 하는 깁스 표집기와는 달리 1000번의 반복만을 통해서도 깁스 표집기의 결과와 유사한 결과를 나타냄을 확인할 수 있었다. 역 베이즈 공식 표집기의 결과는 정확 표집기(exact sampler)로 부터의 결과이므

표 **3.5.** 깁스 표집기 반복 횟수에 따른 Kullback-Leibler divergence 값

iteration (J)	1000	10000	100000
KL divergence	0.006770828	−0.001573642	0.0003379718

로 수렴여부를 확인할 필요가 없으며, 동시에 쿨백 라이블러 발산(Kullback-Leibler divergence) 평가기법을 사용하여 깁스 표집기의 수렴여부를 확인하는 대안적인 방법에 활용될 수 있다. 구체적으로, 깁스 표집기와 역 베이즈 공식 표집기로 얻어진 사후표본을 이용하여 두 표본들 사이의 무질서도(entropy)가 얼마나 일치하는지를 확인하는 방법이 Kullback-Leibler 발산 평가 기법이며, 두 사후표본의 무질서도의 비(ratio)에 로그를 취한 값이 쿨백 라이블러 발산으로 정의되므로, 무질서도가 비슷해질수록 그 비율은 1에 가까우며 쿨백 라이블러 발산 값은 영에 가까운 값을 갖게 된다는 사실을 이용한다. 이를 바탕으로 Kullback-Leibler 표본 통계량을 계산하기위한 방식은 다음과 같다. 먼저, 깁스 표집기와 역 베이즈 공식 표집기에서 각각 추출된 표본들이 식 (3.2)와 같이 주어졌다고 하자.

$$\left\{\lambda_{Gibbs}^{(j)}, \phi_{Gibbs}^{(j)}, z_{Gibbs}^{(j)}\right\}_{j=1}^{J}, \qquad \left\{\lambda_{IBF}^{(k)}, \phi_{IBF}^{(k)}, z_{IBF}^{(k)}\right\}_{k=1}^{K} \tag{3.2}$$

식 (3.2)의 결과들을 바탕으로 각각의 주변 사후분포(marginal posterior distribution)를 식 (3.3)과 (3.4)와 같이 근사적으로 계산한다.

$$p^{Gibbs}(\lambda|y_{obs}) = \int p(\lambda|y_{obs}, z)p^{Gibbs}(z|y_{obs})dz \approx \frac{1}{J}\sum_{j=1}^{J} p\left(\lambda|y_{obs}, z_{Gibbs}^{(j)}\right), \tag{3.3}$$

$$p^{IBF}(\lambda|y_{obs}) = \int p(\lambda|y_{obs}, z)p^{IBF}(z|y_{obs})dz \approx \frac{1}{K}\sum_{k=1}^{K} p\left(\lambda|y_{obs}, z_{IBF}^{(k)}\right). \tag{3.4}$$

따라서 $p^{IBF}(\lambda|y_{obs})$와 $p^{Gibbs}(\lambda|y_{obs})$ 간의 쿨백 라이블러 발산은 식 (3.5)와 같이 정의되며 식 (3.3)과 (3.4)를 이용하여 식 (3.6)과 같은 근사식을 얻는다.

$$\mathrm{KL}\left(p^{IBF}(\lambda|y_{obs}), p^{Gibbs}(\lambda|y_{obs})\right) = \int p^{IBF}(\lambda|y_{obs}) \log \frac{p^{IBF}(\lambda|y_{obs})}{p^{Gibbs}(\lambda|y_{obs})} d\lambda \tag{3.5}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} \log \frac{p^{IBF}(\lambda^{(l)}|y_{obs})}{p^{Gibbs}(\lambda^{(l)}|y_{obs})}. \tag{3.6}$$

식 (3.5)에서 $\lambda^{(1)}, \ldots, \lambda^{(L)}$은 $p^{IBF}(\lambda|y_{obs})$로부터 추출된 랜덤표본을 의미한다. 식 (3.3)–(3.5)의 근사식에서는 서로 다른 상수 J, K, L이 등장하는데 이 중 깁스 표집기의 반복 값인 J가 주 관심대상인 반복수이며 깁스 표집기가 수렴하였다면 값이 증가함에 따라서 쿨백 라이블러 발산 값이 0에 수렴할 것으로 기대할 수 있다 (Tan 등, 2010). 이러한 개념을 바탕으로 Trajan 자료에 적용한 결과가 표 3.5에 요약되어 있다. 깁스 표집기의 반복 횟수가 커질수록 쿨백 라이블러 발산값이 0에 가까운 값으로 나타남을 알 수 있는데 이는 깁스 표집기를 통한 사후 표본과 역 베이즈 공식 표집기로부터의 사후 표본 사이의 무질서도(entropy)가 반복횟수가 커짐에 따라 점점 같아진다는 뜻이다. 즉, 역 베이즈 공식 표집기는 수렴성 여부를 점검할 필요가 없는 정확 표집기이므로 깁스 표집기의 결과물들과 무질서도가 같아진다는 것은 깁스 표집기의 수렴을 의미한다고 이해할 수 있으며 깁스 표집기 사후 표본의 수렴여부를 확인하는 대안적인 방식으로 역 베이즈 공식 표집기가 활용될 수 있다는 것을 의미한다. 역으로, 수렴여부가 확인된 깁스 표집기 사후표본은 역 베이즈 공식 표집기에서의 표본추출이 제대로 이루어졌는가를 확인하는데 같은 방식으로 사용될 수도 있을 것이다. 아울러, 2절에서의 언급과, 표 3.3의 결과에서처럼 역 베이즈 공식 표집기를 사용하는 경우에는, 사후분포의 최빈값을 주어진 모수값 (ϕ_0, λ_0)으로 사용함으로써 q_k의 안정성(stability), 즉 작은 변동성(variation, $\mathrm{Var}(q_k)$)을 확보하도록 한다.

표 **3.6.** 광주기 16시간 자료 적합을 위한 포아송 모형과 ZIP 모형간의 비교

	Poisson	ZIP
DIC	842.6	381.8
베이지안 χ^2 적합도	3265.148	56.73583

3.3. 베이지안 모형 선택 및 계층적 베이지안 모형: Rodrigues (2003) 보완

3.1절에서 언급된 바와 같이 Rodrigues (2003)에서는 계수형 자료인 Trajan 자료를 분석하기 위한 모형으로서 포아송 모형과 영 과잉 포아송 모형을 고려했으며, 표 3.2의 적합 빈도(Fitted frequency)를 통해서 볼 때도 포아송 모형에 비해서 영 과잉 포아송 모형이 잘 적합함을 알 수 있으며, 두 모형 간의 선택 기준으로 카이제곱 적합 값을 제공하였고, 특히 광주기(photoperiod) 16시간의 경우는 영 과잉 모형이 선호됨을 알 수 있었다. 다만, Rodrigues (2003)의 결과에는 표 3.2의 카이제곱 적합결과가 어떤 식으로 계산 되었는지에 대한 언급이 없으며, 몇 가지 미흡한 결과들이 보이기 때문에 다음과 같은 방식으로 이러한 점들을 보완하고자 한다. 첫 번째로는, 표 3.2와 같은 카이제곱 적합도를 얻기 위하여 베이지안 관점의 카이제곱 적합도 검정을 고려한다. 즉, 전통적인 카이제곱 적합도에 대응하는 개념인 베이지안 카이제곱 적합도 검정은 사후분포로 부터 추출된 모수의 사후표본 값을 바탕으로 계산되는 카이제곱 적합도 검정을 의미하며, 일반적인 정의와 그에 따른 이론적인 결과들은 Johnson (2004)을 참조할 수 있다. 두 번째로는 Trajan data에 대하여 영 과잉 포아송 모형이 적합한지에 대한 여부를 카이제곱 적합외에 대안으로서 DIC(Deviance Information Criterion)를 사용하여 확인해본다. 세 번째로는 Rodrigues (2003)의 자료 분석에서는 고려하지 않았던 BAP 농도에 따른 차이를 모형화한다. 구체적으로는 광주기 16시간에 따른 뿌리의 수 전체 합에 대한 모형 대신에 네 가지 BAP 농도 (2.2, 4.4, 8.8, 17.2)에 따른 뿌리의 수에 대하여 베이지안 영 과잉 모형을 적용한다. 또한 이 경우, 네 가지 농도 차에 따른 서로 다른 모수를 적합하기 위하여 계층적 베이지안 모형(hierarchical Bayesian model)의 사용도 추가적으로 고려하고 이를 바탕으로 사후 예측 점검을 통한 카이제곱 적합도를 계산해본다. 이러한 점들을 보완한 광주기 16시간에 대한 추가적인 자료 분석 결과와 그에 따른 논의는 다음과 같다. Trajan 자료에 대하여 포아송 모형과 영 과잉 모형 사이에 어떤 모형이 더 적합하지를 위해서는 기본적으로 다음과 같은 가설 검정 또는 모형 선택을 고려한다.

$$\begin{aligned} & H_0 : \phi = 0, \quad vs. \quad H_1 : \phi > 0 \\ \Leftrightarrow\ & M_0 : Y_{obs} \sim \text{Poisson}(\lambda), \quad vs. \quad M_1 : Y_{obs} \sim \text{ZIP}(\phi, \lambda) \end{aligned} \tag{3.7}$$

Rodrigues (2003)에서는 식 (3.7)에 주어진 모형 선택을 위하여 카이제곱 적합도를 사용했고 빈도론적 관점을 바탕으로, Broek (1995)에서는 스코어 검정(score test)을 제안했고, Ridout 등 (2001)에서는 Trajan 자료에 대하여 스코어 검정을 이용하였다. 식 (3.7)에 대한 베이지안 모형 선택을 위해서는 베이지안 자료 분석에서 실제적으로 자주 사용되는 기준 중의 하나인 DIC(Deviance Information criterion, Spiegelhalter 등, 2002)를 이용한다. DIC를 기준으로 모형을 선택하는 경우에는 더 작은 값의 DIC를 갖는 모형이 선호되며, 특히 WinBUGS에서는 이러한 DIC 값을 자동적으로 제공하며 (Ntzoufras, 2009), 실제 자료 분석에서 유용하게 사용된다 (Ghosh 등, 2006; Johnson, 2004). 광주기 16시간 자료를 바탕으로한 식 (3.7)의 검정 및 선택을 위한 DIC를 계산하면 표 3.6과 같다.

표 3.6의 결과에서 확인할 수 있는 것처럼 DIC를 통한 모형 비교에서는, 일반적인 포아송(Poisson) 모형 적합시보다 영 과잉 포아송(ZIP) 적합을 했을 때 더 낮은 DIC 값을 갖기 때문에 광주기 16시간 자료에 대해서는 영 과잉 포아송(ZIP) 모형적합이 더 타당해 보인다. 이러한 결과는 표 3.2에 나와 있는 실제 관측값이나 Rodrigues (2003)의 카이제곱 적합 값이 보여주는 결과와 같은 결론을 보여준다. 그러

표 **3.7.** 사후 표본 요약: 광주기 16시간/4가지 BAP 농도

사후표본	mean	s.d	2.5	25	median	75	97.5
λ_1	6.496	0.673	5.276	6.033	6.472	6.914	7.910
ϕ_1	0.491	0.077	0.342	0.437	0.489	0.544	0.645
λ_2	5.839	0.638	4.665	5.392	5.817	6.268	7.155
ϕ_2	0.512	0.078	0.364	0.458	0.513	0.564	0.663
λ_3	5.185	0.544	4.190	4.811	5.172	5.555	6.295
ϕ_3	0.422	0.077	0.276	0.370	0.422	0.474	0.572
λ_4	4.639	0.472	3.726	4.314	4.629	4.948	5.613
ϕ_4	0.471	0.070	0.337	0.423	0.469	0.518	0.610

나, 표 3.2의 카이제곱 적합 값의 결과에서 알 수 있듯이 포아송 모형의 경우에 비해서 상대적으로 ZIP 모형의 카이제곱 적합 값이 작은 것이지만 실제로 자유도 10 $(13-2-1)$을 따르는 p-값을 계산해보면 거의 0에 가까운 값을 보여주며, 두 모형 다 적합하지 않음을 알 수 있다. 이러한 문제점을 보완하기 위하여, 우리가 고려해보고자 하는 방식은 광주기 16시간 자료를 BAP농도에 따라서 구분하여, 각각 영 과잉 포아송 모형을 적용하는 것이다. 구체적으로는 광주기 16시간에 따른 뿌리의 수 전체 합에 대한 모형 대신에 네 가지 BAP 농도 (2.2, 4.4, 8.8, 17.2) 에 따른 뿌리의 수에 대하여 각각 베이지안 영 과잉 포아송 모형을 적용한다. 또한 이 경우, 네 가지 농도 차에 따른 서로 다른 모수를 적합하기 위하여 식 (3.8)과 같은 계층적 베이지안 모형(hierarchical Bayesian model)을 사용하고 이를 바탕으로 사후 예측 점검을 통한 카이제곱 적합도를 계산해본다.

$$
\begin{aligned}
Y_{ij} &\sim \text{ZID}(\phi_j, \lambda_j), \quad i = 1, \ldots, n_j \\
\phi_j &\sim \text{Beta}(a, b), \\
\eta &\sim \text{Beta}(1, 1), \\
\psi &\sim \text{Gamma}(0.1, 0.1), \quad a = \eta \cdot \psi,\ b = (1-\eta) \cdot \psi, \\
\lambda_j &\sim \text{Gamma}(c, d), \quad j = 1, 2, 3, 4, \\
c &\sim \text{Exp}(1), \\
d &\sim \text{Gamma}(0.1, 1).
\end{aligned} \tag{3.8}
$$

식 (3.8)에서는 네 가지 서로 다른 BAP 농도(2.2, 4.4, 8.8, 17.2)를 나타내기 위하여 첨자 $j = 1, 2, 3, 4$를 사용하며, 계층적 베이지안 구조를 설명하기 위해서 초모수(hyperparameter)에 대한 사전분포를 설정한다. 초모수 사전분포(hyperprior)를 설정하는데 있어서, ϕ에 대해서는 두 모수 a, b와 베타확률변수의 평균 및 분산과의 관계를 통해서 초모수 사전분포가 할당되었고, 포아송 모수 λ의 초모수 c와 d에 대해서는 각각 지수분포와 감마분포를 초모수 사전분포로 설정하였다. 이와 같은 초모수 사전분포는 이항모형에 대한 계층적 모형과 포아송 모형에 대한 계층적 베이지안 모형에서 많이 사용되는 초모수 사전분포이다 (George 등, 1993; Spiegelhalter 등, 2003, Pump data; Christensen 등, 2011; 등).

계층적 베이지안 모형 적합 결과는 표 3.7에 요약되어 있다. 표 3.7의 결과를 보면 서로 다른 농도에 따른 뿌리의 개수를 설명하는 영 과잉 포아송 모형의 모수 (λ, ϕ)에 차이가 있음을 알 수 있으며, 실제 관측 값과 사후 예측을 통한 적합 값을 통해서 볼 때도 서로 다름을 알 수 있었다. 특히 카이제곱 적합값과 그에 따른 p-값을 살펴보면 표 3.8을 통해 볼 때 BAP 농도 2.2와 17.2의 자료의 경우 유의하지 않은 p-값으로서, 영 과잉 포아송 모형이 적합함을 나타내고 있으며 BAP 농도 4.4와 8.8은 0에 매우 가까운

표 **3.8.** 사후 예측 적합: 광주기 16시간/4가지 BAP 농도

No. of roots	$\text{Obs}^{2.2}$	$E_{ZIP}^{2.2}$	$\text{Obs}^{4.4}$	$E_{ZIP}^{4.4}$	$\text{Obs}^{8.8}$	$E_{ZIP}^{8.8}$	$\text{Obs}^{17.6}$	$E_{ZIP}^{17.6}$
0	15	14.75	16	15.47	12	12.78	19	19.08
1	0	0.17	2	0.28	3	0.55	2	1.01
2	2	0.53	1	0.78	2	1.36	2	2.26
3	2	1.10	1	1.45	1	2.28	4	3.40
4	1	1.73	2	2.06	2	2.88	3	3.87
5	2	2.19	1	2.35	2	2.95	1	3.57
6	1	2.33	2	2.27	3	2.55	4	2.77
7	0	2.16	0	1.89	1	1.90	3	1.86
8	1	1.76	1	1.40	0	1.26	0	1.10
9	3	1.29	0	0.93	2	0.74	2	0.59
10	1	0.86	3	0.56	0	0.40	0	0.28
11	1	0.53	0	0.31	1	0.20	0	0.13
12	1	0.30	1	0.16	1	0.09	0	0.05
χ^2		19.01		44.08		45.61		13.23
p-값		0.09		0.00		0.00		0.30

p-값, 즉 유의한 값을 가지며, 영 과잉 포아송 모형이 적합하지 않음을 나타내고 있다. 다만, BAP 농도 2.2에서의 p-값은 0.09로서 다소 모호한 결과를 나타낸다. 이러한 결과들을 Rodrigues (2003)에서 보여주고 있는 광주기 16시간 전체를 모형화한 결과에서와 같은 영 과잉 포아송 모형의 부적합 결과를 보완해주는 결과라고 할 수 있다.

4. 결론

본 논문에서는 영 과잉 포아송 모형의 베이지안 분석을 위하여 사후 표본 추출 시 고려해 볼 수 있는 두 가지 방법을 고려하였다. 이를 위하여 MCMC 방법을 통한 반복적 추출 방법과 역 베이즈 공식 표집기에 의한 정확 표집법 두 가지를 고려하고 실제 응용문제에서의 자료에 대해 분석하고 그 결과를 논의하였다. 구체적으로는 깁스 표집기(Gibbs sampler)를 이용하는 MCMC 방법을 설명하고 깁스 표집기로부터 추출된 사후표본의 수렴성 여부를 확인하는 여러 가지 기준을 논의하였다. 아울러 역 베이즈 공식 표집기를 통한 사후표본 추출법을 설명하고 이를 이용하여 쿨백-라이블러 발산(Kullback-Leibler divergence)기준을 통한 깁스 표집기 사후표본의 수렴성 여부를 확인하는 방식에 대해서도 설명하였다. 이러한 결과들을 바탕으로 Trajan이라는 사과 품종의 발아에 관한 실제 자료(Trajan data, Marin 등, 1993)를 분석하고 기존의 자료 분석 결과 (Rodrigues, 2003)에서 고려하지 않았던 부분을 비교하고 보완하였다. 본 논문에서 고려한 방식인 BAP 농도에 따른 세부적인 분석이 더 향상된 결과를 보였으며 기존의 결과에서 보여주었던 베이지안 영 과잉 포아송 모형의 부적합성을 다소 완화할 수 있었다. 카이제곱 적합도의 유의확률을 통해서 살펴 본 Trajan 자료에 대한 적합도 문제는 대안적으로 영 과잉 음이항(Zero Inflated Negative Binomial; ZINB) 모형과 같은 또 다른 영 과잉 모형을 고려해 볼 수 있는 향후 과제를 남긴다. 실제로 빈도론적인 관점에서 Ridout 등 (2001)에서는 스코어 검정을 통해서 Trajan 자료를 사용하여 ZIP과 ZINB간의 모형 비교를 실시하였다. 본 논문에서 고려하는 방식을 바탕으로 ZIP과 ZINB모형간의 베이지안 모형 적합과 그에 따른 모형 선택을 고려하는 것은 흥미로운 향후 과제라고 할 수 있다. 또한 본 논문에서 소개한 역 베이즈 공식 표집기를 ZINB모형과 선형혼합(linear mixed)모형으로 확장하고 역 베이즈 공식 표집기 사후 표본을 이용한 베이즈 인자(Bayes factor)계산과 이를 통한 모형 선택 방법론을 개발하고, 추가적으로 영 과잉 포아송 회귀 모형이나 영 과잉 음이항 회

귀 모형을 본 논문에서 고찰한 결과들을 바탕으로 베이지안 관점에서 고찰하고 자료 분석에 활용하는 것이 필요하다고 여겨진다. 이러한 추가 연구들은 본 저자들에 의해 현재 진행 중이거나 향후 이루어질 연구방향이라고 하겠다.

부록

본 논문에서 사용한 기본적인 ZIP 모형 (3.1)의 구현을 위한 WinBUGS 코드는 다음과 같다.

```
model{
     for(i in 1:n){
                    y[i]~dpois(mu[i])
                    mu[i]<-(1-u[i])*lambda
                    u[i]~dbern(phi)
                              }
 lambda~dgamma(0.001,0.001)
 phi~dbeta(0.5,0.5)
 }
```

참고문헌

이지호 (2011). <영과잉 포아송 분포에 대한 베이지안 방법론 고찰>, 고려대학교 통계학과 석사학위논문.

Albert, J. (2007). *Bayesian Computation with R*, Use R! Springer, New York.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669–679.

Angers, J. F. and Biswas, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model, *Computational Statistics & Data Analysis*, **42**, 37–46.

Broek, J. van den (1995). A score test for zero inflation in a poisson distribution, *Biometrics*, **51**, 738–743.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2011). Bayesian ideas and data analysis, *Texts in Statistical Science Series*, CRC Press, Boca Raton, FL. An introduction for scientists and statisticians.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society: Series B*, **56**, 363–375.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–472.

George, E. I., Makov, U. E. and Smith, A. F. M. (1993). Conjugate likelihood distributions, *Scandinavian Journal of Statistics*, **20**, 147–156.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, In *Bayesian Statistics*, 4 (Peñíscola, 1991), 169–193. Oxford University Press, New York.

Ghosh, S. K., Mukhopadhyay, P. and Lu, J. C. (2006). Bayesian analysis of zero-inflated regression models, *Journal of Statistical Planning and Inference*, **136**, 1360–1375.

Gómez-Rubio, V. and López-Quílez, A. (2010). Statistical methods for the geographical analysis of rare diseases, *Advances in Experimental Medicine and Biology*, **686**, 151–171.

Heilbron, D. C., Jewell, N. P., Hauck, W. W., Fusaro, R. E., Kalbfleisch, J. D., Neuhaus, J. M. and Ashby, M. A. (1989). An annotated bibliography of quantitative methodology relating to the AIDS epidemic, *Statistical Science*, **4**, 264–281.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions*, third edition, Wiley Series in Probability and Statistics, John Wiley & Sons, New Jersey.

Johnson, V. E. (2004). A Bayesian χ^2 test for goodness-of-fit, *Annals of Statistics*, **32**, 2361–2384.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Marin, J., Jones, O. and Hadlow, W. (1993). Micropropagation of columnar apple trees, *Journal of Horticultural Science*, **68**, 289–297.

Ntzoufras, I. (2009). *Bayesian Modeling using WinBUGS*, Wiley, Hoboken, New Jersey.

Raftery, A. E. and Lewis, S. M. (1992). [Practical markov chain monte carlo]: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo, *Statistical Science*, **7**, 493–497.

Ridout, M., Hinde, J. and Demétrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, *Biometrics*, **57**, 219–223.

Rodrigues, J. (2003). Bayesian analysis of zero-inflated distributions, *Communications in Statistics - Theory and Methods*, **32**, 281–289.

Rubin, D. B. and Gelman, A. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–472.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, K. (2003). *WinBUGS User Manual*, MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society - Series B (Statistical Methodology)*, **64**, 583–639.

Tan, M. T., Tian, G. L. and Ng, K. W. (2003). A noniterative sampling method for computing posteriors in the structure of EM-type algorithms, *Statistica Sinica*, **13**, 625–639.

Tan, M. T., Tian, G. L. and Ng, K. W. (2010). *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*, Chapman & Hall/CRC CRC Press, Florida.

Tian, G. L., Tan, M. and Ng, K. W. (2007). An exact non-iterative sampling procedure for discrete missing data problems, *Statistica Neerlandica*, **61**, 232–242.

Ugarte, M. D. and Militino, A. F. (2004). Testing for poisson zero inflation in disease mapping, *Biometrical Journal*, **46**, 526–539.

Yip, P. (1988). Inference about the mean of a Poisson distribution in the presence of a nuisance parameter, *Australian Journal of Statistics*, **30**, 299–306.

Bayesian Approaches to Zero Inflated Poisson Model

Jiho Lee[1] · Taeryon Choi[2] · YoonSung Woo[3]

[1]Department of Statistics, Korea University; [2]Department of Statistics, Korea University
[3]Department of Statistics, Korea University

(Received February 2011; accepted June 2011)

Abstract

In this paper, we consider Bayesian approaches to zero inflated Poisson model, one of the popular models to analyze zero inflated count data. To generate posterior samples, we deal with a Markov Chain Monte Carlo method using a Gibbs sampler and an exact sampling method using an Inverse Bayes Formula(IBF). Posterior sampling algorithms using two methods are compared, and a convergence checking for a Gibbs sampler is discussed, in particular using posterior samples from IBF sampling. Based on these sampling methods, a real data analysis is performed for Trajan data (Marin *et al.*, 1993) and our results are compared with existing Trajan data analysis. We also discuss model selection issues for Trajan data between the Poisson model and zero inflated Poisson model using various criteria. In addition, we complement the previous work by Rodrigues (2003) via further data analysis using a hierarchical Bayesian model.

Keywords: Gibbs sampler, Inverse Bayes Formula, Bayesian χ^2 goodness of fit, DIC, hierarchical Bayesian model.

This research was supported by the Basic Science Rsearch Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0010422). In addition, this research is an extension of the first author, Jiho Lee's Master thesis.

[2]Corresponding author: Associate Professor, Department of Statistics, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-701, Korea. E-mail: trchoi@korea.ac.kr