응용통계연구 제 14권 1호, 2001년, pp. 161-175

준모수적 계층적 선택모형에 대한 베이지안 방법

정윤식[1] 장정훈 [2]

요 약

메타분석(Meta-analysis)은 서로 독립적으로 연구되어진 결과들을 전체적인 하나의 결과로 도출하기 위해 사용되어지는 통계적 방법이다. 이러한 통계적 방법을 설명할 모형으로는 선택모형(selection model)을 포함한 계층적 모형(hierarchical model)을 사용하며, 이러한 모형들은 베이지안 메타분석에 유용한 것으로 알려져 있다. 그러나, 메타분석의 자료들은 일반적으로 출판편의(publication bias)를 갖고 있으므로 이를 극복하고자 가중함수(weight function)를 이용하여 분포함수를 새롭게 정의하여 사용한다. 최근에 Silliman(1997)은 계층적 모형(hierarchical model)에 가중함수를 첨부한 계층적 선택모형(hierarchical selection model)을 정의하고 모수적 베이지안 방법을 제시하였다. 본 연구에서는 미관측된 연구효과에 디리슈레 과정 사전분포(Dirichlet process prior) 를 적용한 준모수적 계층적 선택모형(semiparametric hierarchical selection models)을 소개한다. 여기서 제시된 준모수적 계층적 선택모형을 베이지안 방법으로 추정하기 위하여 마코프 연쇄 몬테칼로(Markov chain Monte Carlo) 방법을 이용한다. 제시된 방법을 적용하기 위하여 실제 자료(Johnson, 1993)인 충치를 예방하기 위한 두 가지의 예방약의 효과에 대한 차이를 비교하기 위해 얻어진 12개의 연구를 이용하여 메타분석을 한다.

주요용어: 베이지안 메타분석, 임상정보 사전분포, 디리슈레 과정 사전분포, 깁스 샘플러, 계층적 선택모형, 메트로폴리스-해스팅 알고리즘, 혼합 디리슈레 사전분포, 가중함수.

1. 서론

메타분석은 서로 독립적으로 연구되어진 결과들을 종합하여 전체적인 하나의 결과를 도출하기 위한 통계적 방법으로 특히 의학 분야에서 어떤 약의 효과라든지 새로운 연구에 대한 계획 등에 널리 쓰이고 있다. 이러한 메타분석에는 두 가지의 중요한 문제점이 있다. 하나는 메타분석에 포함되는 연구들은 각기 다른 시점, 장소에서 서로 독립적으로 연구된 결과이므로 각각의 연구효과가 이질적(heterogeneous)이라는 것으로 대개 임의효과 모형(random effect model)이나 계층적 모형(hierarchical model)에 의해 설명된다.(Morris와 Normand, 1992). 메타분석에 계층적 모형을 적용하는 것은 관심의 대상이 되는 모집단으로부터 추출된 표본으로부터 전체적인 평균 μ를 추정하는 데 주로 목적을 두기 때문에 오래 전부터 중요시 되어 왔다(Hedges와 Olkin, 1985). 또한 베이지안 입장에서 이 모형은 관심있는 모수에 대하여 여러 가지의 초월 사전분포 (hyperprior)를 고려하는 것으로 이러한

1) (609-735) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과, 부교수
E-mail: yschung@hyowon.cc.pusan.ac.kr
2) (609-735) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과, 박사과정

초월 사전분포의 변화에 따른 로버스트성을 쉽게 조사할 수 있는 장점을 가지고 있으므로 이러한 계층적 모형은 메타분석에서 가장 자주 쓰이고 유용한 모형이라고 할 수 있다. 최근에 Larose와 Dey(1997), Chung과 Jeong(2000)은 그룹화된 임의 효과 모형에 로짓모형과 프랏빗모형을 이용하여 각각 연구하였다. 또 다른 문제점은 출판편의(publication bias)가 존재하는 것이다. 예를 들면 연구자들이 구할수 있는 자료들은 거의가 유의한 결과를 보여주는 연구들로써 이들은 모두 출판물로 나타난 것들이다. 즉, 다시 말해서 유의한 결과를 보여주는 연구는 그렇지 못한 연구에 비해 출판될 확률이 높다는 것으로 자료를 수집할 때 통계적으로 유의한 결과를 보여주는 연구가 메타분석에 포함될 가능성이 커지는 것을 의미하므로 이 때 편의(bias)가 발생한다. 이와 같이 일정한 수준을 넘어선 것들만을 얻게되므로 이를 출판편의라 한다. 그러므로, 출판편의가 존재할 때 이러한 편의를 보정하기 위하여 일반적으로 가중함수(weight function)가 사용된다 (Larose와 Dey, 1996).

Silliman(1997)은 계층적 모형에 가중함수를 포함시킨 계층적 선택모형(hierarchical selection models)을 소개했다. 이 모형은 각각의 연구효과의 이질성과 출판편의를 동시에 설명하고 각 미관측된 연구효과 α_i가 평균 μ과 분산 σ_α^2을 갖는 정규분포를 따르는 임의효과로 간주함으로서 이 모형은 모수적 임의효과 모형과 유사하다. 그러나 이러한 임의효과에 완전히 모수적인 접근(fully parametric approach)을 함으로서 분포에 대한 가정이 잘못되었을 때는 결과가 잘못 도출될 수 있는 문제점을 가지고 있다. 그러므로 본 연구에서는 이러한 문제점을 해결하기 위해 미관측된 연구효과의 분포에 디리슈레 과정 사전분포(Dirichlet process prior)를 적용한다. 즉, 미 관측된 연구효과들의 분포 G가 현 자료를 분석하는 시점에서 알려져 있지 않다고 가정한다. 이는 G를 랜덤확률분포로 가정한 것이다. 그러므로 미관측 효과의 분포 G가 랜덤이므로 디리슈레 과정(Dirichlet process)을 따른다 한다. 이 때 이를 준모수적 계층적 선택모형(semiparametric hierarchical selection model)이라 정의한다. 제시된 준모수적 계층적 선택모형은 임의효과 분포의 잘못된 가정을 어느 정도 보정할 수 있다. 이러한 디리슈레 과정 사전분포는 Ferguson(1973)에 의해 소개되었고 Antoniak(1974)이 이것을 좀 더 보완하여 혼합 디리슈레 과정 (mixture of Dirichlet process)을 소개하였다. 이러한 비모수적 방법은 관심의 대상이 되는 모수 θ의 분포에 대한 확신이 없을 때 사용한다. 특히 디리슈레 과정 사전분포는 연구자가 가장 근접하게 추측하는 분포 G_0를 이 과정의 평균으로 하고, G_0에 대한 믿음의 정도 M을 갖는 확률과정이다. 만약 M의 값이 커져가면 모수 θ의 분포는 G_0에 가까이 가며, M의 값이 작아지면 모수 θ의 분포는 G_0에서 멀어짐을 알 수 있다. 여기서 G_0를 기저 사전분포함수(baseline prior distribution)라 한다. 그들을 이용한 사후분포를 계산하는 과정이 너무 복잡하여 거의 사용하지 못하였다. 그러나 90년대에 들어와 깁스 샘플러(Gelfand와 Smith, 1990)의 등장 이후 계산과정을 상대적으로 많이 완화시켰으므로, 현재 비모수적 베이지안 접근에서 가장 많이 쓰이고 있다. 자세한 계산과정은 West, Müller와 Escobar(1994)을 참조하자.

따라서 본 논문에서는 앞에서 소개된 두 가지 문제점을 해결할 수 있는 준모수적 계층적 선택모형을 제시하여 베이지안 방법으로 해석한다. 이때 여러 형태의 기저 분포함수들과 가중함수들을 제시하여 그 들의 민감도를 조사한다. 또한 M의 크기에 따라 원하는 추정치의 값에 어느 정도의 영향이 있는지 조사한다. 마지막으로, 제시된 방법의 타당성을 보

이기 위해 두 가지 충치예방약의 효과를 비교한 12개의 연구에 대한 메타분석을 베이지안 입장에서 분석할 것이다.

이 논문은 다음과 같이 구성되어 있다. 2장에서는 베이지안 메타분석을 위한 준모수적 계층적 선택모형을 어떻게 구축하는 가에 대한 설명을 하고 3장에서는 정규분포함수를 기저분포(baseline distribution)로 한 계산적 문제를 자세하게 다룰 것이다. 4장에서는 Johnson(1993)의 자료를 설명하고 이 자료로 우리가 제시한 방법을 적용한다. 마지막으로 5장에서 결과에 대한 설명과 앞으로의 과제에 대하여 논한다.

2. 모형

이 장에서는 Silliman(1997)의 계층적 선택모형을 소개하고 그것을 Ferguson(1973)에 의해 제시된 디리슈레 과정 사전분포를 이용한 준모수적 계층적 선택모형으로 확장한다.

관심이 있는 정규모집단으로부터 확률표본을 얻을 수 있다고 가정하자. 이 때 Morris와 Normand(1992)가 고려한 계층적 모형은 다음과 같다. $i = 1, \ldots, n$에 대하여

$$\begin{aligned} Y_i|\alpha_i, \sigma_i &\sim N(\alpha_i, \sigma_i^2), \\ \alpha_i|\mu, \sigma_\alpha^2 &\sim N(\mu, \sigma_\alpha^2), \\ \mu &\sim N(a, b) \end{aligned}$$

와

$$\sigma_\alpha^2 \sim IG(c, d), \tag{2.1}$$

여기서 $IG(c,d)$는 모양모수(shape parameter) c와 척도모수(scale parameter) d를 가지는 역 감마 분포(Inverse Gamma distribution)를 나타낸다. 메타분석의 관점에서 Y_i는 관측된 연구효과, α_i는 미관측된 실제 연구효과 , σ_i^2는 연구내 분산, μ는 전체적인 연구효과 그리고 σ_α^2는 연구간 분산이라고 해석할 수 있다. 일반적으로 σ_i는 표준오차로 추정되므로 본 연구에서도 이것을 고정시킨다. 이에 대한 자세한 내용은 Johnson(1993)을 참조하고, (2.1)로부터 μ와 σ_α^2 은 공액사전분포(conjugate prior)가 가정되었다는 것을 알 수 있다.

메타분석에서 모형 (2.1)은 그 자체가 통계적으로 유의한 쪽으로 편의된 출판 연구(published research)의 하나이기 때문에 자주 쓰이는 모형은 아니다. 그러한 문제를 해결하기 위해서 Larose와 Dey(1996)는 가중함수들을 제시하고 베이즈 인자를 이용하여 자료에 잘 적합하는 가중함수를 선택하였다. 또한, Silliman(1997)은 (2.1)과 같은 계층적 모형에 가중함수를 포함시킨 계층적 선택모형을 소개하였다. 즉, 임의의 연구 효과 y_i가 관측될 확률은 어떤 음이 아닌 함수 $w(y_i)$와 곱해져서 나타나므로 확률변수 Y_i는 다음의 밀도함수를 가지는 가중분포(weighted distribution)로부터 나온다고 할 수 있다. 즉,

$$f^w(y_i|\alpha_i, \sigma_i) = \frac{w(y_i)\, f(y_i|\alpha_i, \sigma_i)}{C_w}, \tag{2.2}$$

여기서 $f(y_i|\alpha_i,\sigma_i) = N(y_i;\alpha_i,\sigma_i)$이고 $N(y_i;\alpha_i,\sigma_i)$은 확률변수 Y_i의 분포가 평균 μ, 분산 σ_α^2를 갖는 정규분포임을 나타낸다. 그리고

$$C_w = \int w(x)\, f(x|\alpha_i,\sigma_i)dx, \tag{2.3}$$

는 정규화 상수(normalizing constant)이다. 이때, 가중함수 $w(y_i)$의 선택에 대해서는 Larose와 Dey(1996) 와 Silliman(1997)을 참조하라. 따라서 Silliman(1997)의 계층적 선택모형은

$$\begin{aligned} Y_i|\alpha_i,\sigma_i &\sim f^w(y_i|\alpha_i,\sigma_i), \\ \alpha_i|\mu,\sigma_\alpha^2 &\sim N(\mu,\sigma_\alpha^2), \\ \mu &\sim N(a,b), \end{aligned}$$

와

$$\sigma_\alpha^2 \sim IG(c,d), \tag{2.4}$$

와 같이 되고 $f^w(y_i|\alpha_i,\sigma_i)$는 (2.2)에서 정의된 것과 같다.

모형 (2.4)는 모형의 이질성과 출판 편의를 동시에 설명할 수 있지만 미관측된 연구효과에 가정된 분포가 잘못되었을 경우에는 결과를 잘못 도출할 수 있는 문제점을 가지고 있다. 그리하여, 우리는 이러한 잘못된 가정에 보정할 수 있는 디리슈레 과정 사전분포를 이용한 준모수적 계층적 선택모형을 소개한다. 즉,

$$\begin{aligned} Y_i|\alpha_i,\sigma_i &\sim f^w(y_i|\alpha_i,\sigma_i), \\ \alpha_i|G &\sim G, \\ G|M,\mu,\sigma_\alpha^2 &\sim DP(M\cdot G_0(\mu,\sigma_\alpha^2)), \\ \mu,\sigma_\alpha^2 &\sim p_1(\mu,\sigma_\alpha^2) \end{aligned}$$

와

$$M \sim p_2(M), \tag{2.5}$$

여기서 G는 우리가 알지 못하는 모르는 분포함수이고 DP는 디리슈레 과정을 의미한다. 또한 $G|M,\mu,\sigma_\alpha^2 \sim DP(M\cdot G_0(\mu,\sigma_\alpha^2))$인 표현은 랜덤 확률분포 G가 평균이 G_0이고, 정도가 M인 디리슈레 과정을 따름을 나타낸다. 디리슈레 과정에 대한 보다 자세한 내용은 Ferguson(1973)과 Antoniak(1974)을 참조한다. G_0를 앞에서 설명한 *기저 사전분포(baseline prior)*라고 한다. 본 논문에서는 정규분포를 기저 사전분포로 사용한다. 이때 각 분포들이 갖고 있는 평균 μ와 분산 σ_α^2에도 사전분포 함수를 주고, 믿음의 정도인 M에도 사전분포를 주어 베이지안 해석을 한다. 디리슈레과정은 매우 복잡한 계산을 필요로 하므로 우리는 깁스 샘플러(Gelfand and Smith, 1990)와 메트로폴리스-해스팅 알고리즘(Metropolis *et al.*, 1953; Hastings, 1970)과 같은 마코프 연쇄 몬테 칼로(Markov chain Monte Carlo, MCMC) 방법을 이용하여 계산의 복잡성을 해결한다. MCMC를 이용하는 데에는 각 모수에 대한 완전 조건부 분포(full conditional distribution)가 필요하므로 이것은 다음 3장에서 상세히 다루도록 한다.

3. 계산적 문제

이 장에서는 완전 조건부 분포를 포함한 계산적인 문제를 상세하게 다룬다. 이제부터 각괄호(Bracket)는 밀도함수를 나타낸다. 예를 들면 $[X,Y]$, $[X|Y]$와 $[X]$는 각각 결합밀도함수, 조건부 밀도함수와 주변밀도함수를 나타낸다.

각 연구효과의 기저 사전분포가 평균 μ와 분산 σ_α^2를 갖는 정규분포라고 가정하자. 즉 (2.5)에서 $G_0 = N(\mu, \sigma_\alpha^2)$이다. 그러면 $i = 1, \ldots, n$에 대하여

$$\begin{aligned} Y_i|\alpha_i, \sigma_i &\sim [C(\alpha_i)]^{-1} w(y_i) N(\alpha_i, \sigma_i^2), \\ \alpha_i|G &\sim G, \\ G|M, \mu, \sigma_\alpha^2 &\sim DP(M \cdot N(\mu, \sigma_\alpha^2)), \\ \mu &\sim N(a, b), \\ \sigma_\alpha^2 &\sim IG(c, d), \end{aligned}$$

와

$$M \sim p_2(M), \tag{3.1}$$

여기서 $C(\alpha_i)$는 (2.3)과 동일한 정규화 상수이다. 즉,

$$C(\alpha_i) \propto \int w(x) \exp\left(-\frac{(x-\alpha_i)^2}{2\sigma_i^2}\right) dx. \tag{3.2}$$

2장에서도 언급했듯이 μ와 σ_α^2에 대해서는 공액 사전분포를 가정하였으므로 완전 조건부 분포는 다음과 같이 쉽게 구해질 수 있다.

$$[\mu|\mathbf{y}, \alpha, \sigma, \sigma_\alpha^2, M] = N\left(\frac{b\sum \alpha_i + a\sigma_\alpha^2}{bn + \sigma_\alpha^2}, \frac{b\sigma_\alpha^2}{bn + \sigma_\alpha^2}\right), \tag{3.3}$$

와

$$[\sigma_\alpha^2|\mathbf{y}, \alpha, \sigma, \mu, M] = IG\left(\frac{1}{2}n + c, \frac{1}{2}\sum_{i=1}^{n}(\alpha_i - \mu)^2 + d\right), \tag{3.4}$$

여기서 $\mathbf{y} = (y_1, \ldots, y_n)'$, $\alpha = (\alpha_1, \ldots, \alpha_n)'$와 $\sigma = (\sigma_1, \ldots, \sigma_n)'$을 나타낸다.

West, Müller와 Escobar(1994)가 제시한 방법에 의해 α_i의 완전 조건부 분포는 다음과 같은 혼합 디리슈레 과정(mixture of Dirichlet process, Antoniak, 1974)으로 나타난다.

$$[\alpha_i|\mathbf{y}, \sigma, \mu, \sigma_\alpha^2, \alpha_{-i}, M] \propto q_0 g_b(\alpha_i) + \sum_{j \neq i} q_j \delta(\alpha_i|\alpha_j), \tag{3.5}$$

여기서 $q_0 \propto M \int f^w(y_i|\alpha_i, \sigma_i) dG_0(\alpha_i)$, $q_j \propto f^w(y_i|\alpha_j, \sigma_j)$이고 $q_0 + \sum_{j \neq i} q_j = 1$를 만족한다. 그리고 $\alpha_{-i} = (\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n)'$, $g_b(\alpha_i) \propto f^w(y_i|\alpha_i, \sigma_i) dG_0$이다. 즉, $g_b(\alpha_i)$는 α_i의 분포를 디리슈레 과정 대신 정규분포를 사용할 때 생성되어지는 사후분포함수이므로 이를 *기저 사후분포*라고 한다. $\delta(t|u)$는 u에서 확률 1을 가지는 퇴화분포(degenerate

distribution)이다. 여기서 q_0가 해석적으로 계산이 안되므로 몬테칼로 계산(Monte Carlo computation)을 이용한다. 즉,

$$\hat{q}_0 \simeq M\frac{1}{l}\sum_{j=1}^{l} f^w(y_i|\alpha_i^{(j)},\sigma_i), \tag{3.6}$$

와 같이 되고 $\alpha_i^{(j)}$는 $j=1,\ldots,l$에 대하여 $N(\mu,\sigma_\alpha^2)$로부터 나오는 확률변수들이다. 또한 기저 사후분포는

$$g_b(\alpha_i) = [C(\alpha_i)]^{-1}\, N\left(\frac{y_i\sigma_\alpha^2+\mu\sigma_i^2}{\sigma_\alpha^2+\sigma_i^2}, \frac{\sigma_\alpha^2\sigma_i^2}{\sigma_\alpha^2+\sigma_i^2}\right). \tag{3.7}$$

와 같이 된다. 이 때 가중함수를 $w(y)=|y|$로 선택하면 (3.2)는 해석적으로 계산이 되며 그 결과는

$$C(\alpha_i) = \alpha_i\left(1-2\Phi\left(-\frac{\alpha_i}{\sigma_i}\right)\right)+\sqrt{\frac{2}{\pi}}\,\exp\left(-\frac{\alpha_i^2}{2\sigma_i^2}\right), \tag{3.8}$$

와 같이 되고 여기서 $\Phi(\cdot)$는 표준정규분포의 분포함수이다. 또한 가중함수를 $w(y)=|y|^2$로 선택하면

$$C(\alpha_i) = \alpha_i^2+\sigma_i^2, \tag{3.9}$$

가 되어 이것은 α_i와 σ_i가 주어졌을 때 y^2의 기대값과 일치한다. 따라서, α_i는 (3.7)로부터 바로 샘플링이 안되므로 메트로폴리스-해스팅 알고리즘을 사용한다. Chib과 Greenberg(1995)의 방법에 따라 현재 마코프 연쇄의 값이 $\alpha_i^{(t)}=\theta$라면 $\alpha_i^{(t+1)}$의 후보값인 θ^*를 $N\left(\frac{y_i\sigma_\alpha^2+\mu\sigma_i^2}{\sigma_\alpha^2+\sigma_i^2}, \frac{\sigma_\alpha^2\sigma_i^2}{\sigma_\alpha^2+\sigma_i^2}\right)$ 로부터 샘플링하여 채택확률

$$\beta(\theta,\theta^*) = \min\left[\frac{C(\theta)}{C(\theta^*)},1\right], \tag{3.10}$$

로 θ^*를 $\alpha_i^{(t+1)}$값으로 채택한다. 여기서 $C(\cdot)$는 가중함수 $w(y)$의 선택에 따른 (3.8)식이나 (3.9)식과 일치한다. 더욱 자세한 샘플링 알고리즘에 대해서는 West, Müller와 Escobar(1994)를 참조하라.

이제 마지막으로 디리슈레 과정의 정도모수인 M에 대하여 생각해 보기로 하자. k를 α_i의 서로 구별되는 값의 개수라고 하면 만약 M의 값이 작다면 미지의 랜덤한 분포 G는 몇 개의 원소에 집중될 것이고 이것은 k의 값이 작다는 것을 뜻한다. 마찬가지로 M의 값이 크다면 미지의 랜덤한 분포 G는 기저 사전분포인 G_0와 가깝게 된다는 것을 의미하므로 이것은 k의 값이 크게 된다는 것을 뜻한다. 따라서 k의 분포를 고려하여 보면 West(1992) 의 방법으로 $k=1,\ldots,n$에 대하여

$$[k|M,n] = c_n(k)n!M^k\frac{\Gamma(M)}{\Gamma(M+n)}, \tag{3.11}$$

와 같이 되고 여기서 $\Gamma(\cdot)$은 감마함수이고 $c_n(k)=Pr(k|M=1,n)$로서 M에 의존하지 않는 수이다. 이것은 스털링 수(Stirling numbers)의 재귀적 형식(recurrence formula)으로 계

산되어질 수 있다(West, 1992; Escobar and West, 1995). 그러므로 M의 사전분포는 서로 구별되는 개수 k와 관련이 있다고 할 수 있다. 즉,

$$[k|n] = \int [k|M,n][M]dM. \tag{3.12}$$

우리의 모형으로부터 자료는 초기에 다른 모수들과 M은 조건부 독립임을 가정하였으므로 M의 완전 조건부 분포는

$$[M|k,\mathbf{y},\psi] \propto [M|k] \propto [M][k|M], \tag{3.13}$$

로 되고 여기서 $\psi = (\alpha, \sigma, \mu, \sigma_\alpha^2)$를 나타내고 M의 사전분포로서 감마분포와 혼합 감마분포의 두 가지 경우가 고려될 것이다.

먼저 $M \sim Ga(u,v)$이라고 가정하면 (3.11)식의 감마함수는

$$\frac{\Gamma(M)}{\Gamma(M+n)} = \frac{(M+n)\beta(M+1,n)}{M\Gamma(n)}, \tag{3.14}$$

와 같이 쓰여질 수 있고 여기서 $\beta(\cdot,\cdot)$는 베타함수이다. 이 때 (3.13)은 임의의 $k = 1,\ldots,n$에 대하여

$$\begin{aligned}[M|k] &\propto [M]M^{(k-1)}(M+n)\beta(M+1,n) \\ &\propto [M]M^{(k-1)}(M+n)\int_0^1 \eta^M(1-\eta)^{(n-1)}d\eta,\end{aligned} \tag{3.15}$$

와 같이 되어 이것은 다음의 M과 0과 1사이의 연속인 값을 가지는 η의 결합밀도함수로부터의 주변밀도함수와 같은 꼴이 된다. 즉, M과 η의 결합밀도함수는 다음과 같이 표현된다.

$$[M,\eta|k] \propto [M]M^{(k-1)}(M+n)\eta^M(1-\eta)^{(n-1)}. \tag{3.16}$$

따라서 우리는 자료증대 알고리즘(data augmentation technique, Tanner and Wong, 1987)을 적용한다. 먼저 잠재변수 η를 평균 $\frac{M+1}{M+n+1}$인 베타분포로부터 추출하여 새로운 M의 값을 다음의 혼합 감마분포로부터 추출한다.

$$\begin{aligned}[M|\eta,k] &\propto M^{u+k-2}(M+n)\exp[-M(v-\log(\eta))] \\ &\propto M^{u+k-1}\exp[-M(v-\log(\eta))] + nM^{u+k-2}\exp[-M(v-\log(\eta))] \\ &= \pi_\eta Ga(u+k, v-\log(\eta)) + (1-\pi_\eta)Ga(u+k-1, v-\log(\eta)),\end{aligned} \tag{3.17}$$

여기서 가중치 π_η는

$$\frac{\pi_\eta}{1-\pi_\eta} = \frac{u+k-1}{n(v-\log(\eta)} \tag{3.18}$$

와 같이 정의된다.

다음으로 M의 사전분포가 혼합 감마분포라고 가정하여 보자. 즉,

$$M \sim c_1 Ga(u_1, v_1) + c_2 Ga(u_2, v_2), \tag{3.19}$$

이고 $c_1 + c_2 = 1$이다. 위와 비슷한 방법에 의하여 M의 완전 조건부 분포는

$$\begin{aligned} M|\eta, k &\sim c_{i,\eta}\{\pi_{i,\eta} Ga(u_i + k, v_i - \log(\eta)) \\ &+ (1 - \pi_{i,\eta}) Ga(u_i + k - 1, v_i - \log(\eta))\}, \end{aligned} \tag{3.20}$$

로 된다. 또한 가중치는 $i = 1, 2$에 대하여

$$\frac{\pi_{i,\eta}}{1 - \pi_{i,\eta}} = \frac{u_i + k - 1}{n(v_i - \log(\eta))} \tag{3.21}$$

와

$$c_{i,\eta} \propto c_i \frac{\Gamma(u_i + k - 1)}{(v_i - \log(\eta))^{u_i + k - 1}} \left(n + \frac{u_i + k - 1}{(v_i - \log(\eta))} \right), \tag{3.22}$$

이 되고 $\sum_{i=1}^{2} c_{i,\eta} = 1$이다.

4. 예제

Johnson(1993)은 두 가지 충치 예방약- NaF와 SMFP- 의 효과를 비교하기 위해 이미 연구되어진 12개의 연구를 종합했다. 각 연구에서 y_i는 두 가지 약의 평균 차이를 나타낸다. 즉, 이것은 SMFP를 사용한 환자의 충치범위에서 NaF를 사용한 환자의 충치범위를 뺀 값이 되므로 y_i의 값이 양이라면 NaF가 SMFP보다 더 효과적이라는 것을 나타낸다. 또한 각 연구의 표준오차의 추정치 $\hat{\sigma}_i$와 표본 크기 N이 표 4.1에 주어져 있다. 이런 연구로부터 전체적인 효과 $\hat{\mu}$를 추정하는 것이 우리의 관심이다. Johnson(1993)은 가중평균으로 μ의 값을 추정하였고 그 추정치는 .32, 95%신뢰구간은 (.13,.52)로서 NaF가 더 효과적이라는 가설을 뒷받침한다.

Johnson(1993)의 결과가 비록 NaF의 효과가 더 좋다는 것을 나타내지만 그의 메타분석에는 몇 가지 문제점이 있을 수 있다. 특히 그는 전 세계의 문헌에서 찾아낸 12개의 연구를 고려했지만 실제로 찾아낸 연구는 13개였다. 13번째 연구에서는 분산의 추정치를 구할 수가 없어서 메타분석에서 제외시켰고 또한 이 12개의 연구 중에서 5번째 연구를 제외하면 나머지 11개의 연구효과가 양의 값을 나타내고 있다. 이러한 연구로 Johnson의 연구에는 출판편의가 존재한다고 생각하는 것이 타당하고, 또한 Silliman(1997)은 이 자료에 대하여 계층적 선택모형을 적용하였다.

이제 우리는 표 4.1에 주어져 있는 Johnson의 자료에 앞에서 제안한 준모수적 계층적 선택모형을 적용한다. 3장에서 언급한 대로 정규분포가 디리슈레 과정 사전분포의 기저 사전분포로 고려될 것이다. 또한 가중함수는 $w(y) = |y|$와 $w(y) = |y|^2$의 두 개의 함수를 사용한다. 명백히 $w(y) = |y|$는 연구효과가 유의할수록 좀 더 관측이 잘 된다는 것을 의미하고 $w(y) = |y|^2$는 굉장히 많은 출판편의가 존재한다는 것을 뜻한다. 정규화 상수 $C(\alpha_i)$는

표 4.1: Johnson의 자료

	Studies											
	1	2	3	4	5	6	7	8	9	10	11	12
y_i	.86	.33	.47	.50	-.28	.04	.80	.19	.49	.49	.01	.67
$\hat{\sigma}_i$	.57	.56	.35	.25	.54	.28	.78	.13	.28	.24	.08	.17
N	247	326	277	363	343	1490	418	2273	1352	2762	2222	2126

가중함수의 선택에 따라 각각 (3.8)식 또는 (3.9)식과 같이 된다. 관심있는 모수에 대한 사전분포는 $\mu \sim N(0, 0.04)$ 와 $\sigma_\alpha^2 \sim IG(2.0016, 24.96)$로 주어지는데 이는 DerSimonian과 Laird(1986)에 의해 제시된 비 반복 추정치(noniterative estimate)에 기인한다. 이 사전분포는 Johnson(1993)에 의해 임상 정보 사전분포(clinical informative prior)라 불린다. M의 사전분포로는 평균이 1, 5와 30인 세 개의 감마분포와 이 세 감마분포의 혼합감마분포를 고려하고 혼합감마분포가 고려될 때 각 가중치는 서로 같게 둔다. 여기서 $E(M)$의 선택은 기저분포에 대한 믿음의 정도가 작고, 보통이고 크다는 것을 의미한다. 기호의 편이를 위해 다음을 정의하자.

$$\begin{aligned} M_1 &= Ga(2, 0.5), \\ M_2 &= Ga(5, 1), \\ M_3 &= Ga(15, 2), \\ M_4 &= 0.5Ga(2, 0.5) + 0.5Ga(5, 1), \\ M_5 &= 0.5Ga(2, 0.5) + 0.5Ga(15, 2) \end{aligned}$$

와

$$M_6 = 0.5Ga(5, 1) + 0.5Ga(15, 2),$$

예를 들면 $M \sim M_1$라는 것은 M의 사전분포가 $Ga(2, 0.5)$라는 것을 나타낸다. 우리는 M의 사전 분포의 변화에 따라 우리가 관심있는 모수의 변화를 알아볼 것이다. 결과는 표 4.2와 4.3에 주어져 있다. 표 4.4는 M을 고정시켰을 때의 결과를 나타낸다. 메트로폴리스 알고리즘 10,000번이 포함된 깁스 샘플러는 4,000번 수행되었고 최초 2,000번은 번-인(burn-in) 주기로 버렸다. 알고리즘의 수렴성은 S-Plus에서 쉽게 사용할 수 있는 CODA(Best, Cowles와 Vines, 1995)에 의해 Geweke(1992)이 제시한 통계량에 의해 검사되었다. 모든 모수들의 Geweke통계량의 절대값이 1.96보다 작으므로 수렴은 되었다고 할 수 있다. 표 4.2, 4.3과 4.4에서 괄호안의 수는 자료가 주어졌을 때의 μ가 양일 확률을 나타내고 σ_α^2의 값에 있는 E는 10^{-3}을 나타낸다. 즉, $57E$는 0.0057이라는 뜻이 된다. 표 4.4에 있는 $\bar{k}$는 군집(cluster)의 평균 개수를 나타낸다.

Silliman(1997)이 $w(y) = |y|$하에서 계층적 선택모형 (2.4)를 고려하여 μ와 σ_α^2를 추정한 값은 각각 0.19와 0.026이고 $\mu > 0$일 사후확률은 0.99이다. 또한 $w(y) = y^2$하에서 μ와 σ_α^2를

표 4.2: $w(y) = |y|$하에서 μ, σ_α^2와 M의 추정치

	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\hat{M}$
$M \sim M_1$	0.20(0.998)	57E	6.0
$M \sim M_2$	0.20(0.995)	58E	11.9
$M \sim M_3$	0.22(0.996)	57E	41.0
$M \sim M_4$	0.20(0.998)	56E	10.8
$M \sim M_5$	0.20(0.997)	57E	33.2
$M \sim M_6$	0.19(0.995)	58E	17.9

표 4.3: $w(y) = |y|^2$하에서 μ, σ_α^2와 M의 추정치

	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\hat{M}$
$M \sim M_1$	0.20(0.999)	57E	13.0
$M \sim M_2$	0.20(0.999)	57E	20.5
$M \sim M_3$	0.20(0.999)	57E	48.1
$M \sim M_4$	0.20(0.999)	57E	17.3
$M \sim M_5$	0.20(0.999)	57E	35.7
$M \sim M_6$	0.20(0.999)	57E	36.7

추정한 값은 각각 0.15와 0.021이고 $\mu > 0$일 사후확률은 0.98이다.

표 4.2는 가중함수가 $w(y) = |y|$로 주어진 준모수적 계층적 선택모형으로 분석한 $\hat{\mu}$와 $\hat{\sigma}_\alpha^2$ 의 값과 $\mu > 0$일 사후확률을 나타낸다. μ의 추정치는 M_3인 경우를 제외하고 0.19로서 Silliman의 결과와 유사하고 $\mu > 0$일 사후확률은 거의 모든 경우에 있어서 0.999로서 Silliman(1997)의 결과인 0.99와 비교해 볼 때 NaF가 더 효과적이라는 결과는 로버스트하다. 또한 σ_α^2의 추정치는 0.0057로서 Silliman(1997)의 0.026 보다 더 작은 값을 나타낸다.

표 4.3은 가중함수가 $w(y) = y^2$일 때의 결과를 보여준다. $w(y) = |y|$일 때와 유사한 값을 나타내는 것을 알 수 있다. 그러나 μ와 σ_α^2의 추정치는 Silliman(1997)의 결과보다 더 0에 가까워지고 $\mu > 0$의 사후확률은 0.999로서 NaF가 더 효과적이라는 가설을 강력하게 뒷받침한다. 그러므로 표 4.2와 4.3으로부터 $\hat{\mu}$와 $\hat{\sigma}_\alpha^2$의 값은 가중함수와 M의 사전분포의 선택에 대하여 매우 로버스트함을 알 수 있다. 즉, 디리슈레 과정 사전분포를 이용한 우리의 모형은 특별한 가중함수의 선택에 대하여 민감하지 않다고 할 수 있다. 이러한 맥락으로 각 연구 효과에 대한 모수적 접근에 대한 확신이 없을 때 디리슈레 과정 사전분포를 이용한 비모수적 접근이 더 좋다고 할 수 있다. 또한 표 4.2와 4.3에서 M의 사후 추정치는 M의 사전분포의 선택에 따라 어느정도 민감성을 보이고 있다.

위에서 언급한 대로 μ와 σ_α^2의 사후 추정치는 M의 사전분포의 선택에 상관없이 거의

같은 값을 나타내고 있으므로 M을 고정한 상태에서 자료분석을 하여 M을 랜덤으로 간주한 것과 비교하여 보았다. 그림 4.1은 표 4.4에 주어진 M이 고정되었을 때 가중함수 $w(y) = |y|$하에서 μ의 추정된 밀도함수이다. 실선(solid line)은 $M = 1$일 때, 점선(dotted line)은 $M = 5$일 때, 대시선(dashed line)은 $M = 30$일 때이고 긴 대시선(long dashed line)은 $M = \infty = 10^4$일 때이다. 총 12개의 α_i가 있으므로 $\bar{k}$의 값은 12를 넘지 않음을 주의하여라. 표 4.4에서 우리는 $w(y) = |y|^2$일 때의 결과는 $w(y) = |y|$일 때와 거의 같으므로 여기서는 $w(y) = |y|$일 때만 고려한다. 정규성과 가까운 $M = 30$일 때의 평균 군집은 $\bar{k} = 8.4$이고 정규성과 어느정도 떨어진 $M = 5$일 때의 평균 군집은 $\bar{k} = 4.1$ 이 되며 정규성과 거리가 먼

표 4.4: M이 고정되었을 때의 μ, σ_α^2와 $\bar{k}$의 추정치

M	1			5			30			10^4		
	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\bar{k}$	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\bar{k}$	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\bar{k}$	$\hat{\mu}$	$\hat{\sigma}_\alpha^2$	$\bar{k}$
$w(y) = \|y\|$	0.26	57E	1.5	0.26	57E	4.1	0.24	57E	8.4	0.22	57E	11.8
$w(y) = \|y\|^2$	0.28	57E	1.3	0.25	57E	8.2	0.19	57E	9.3	0.19	57E	11.9

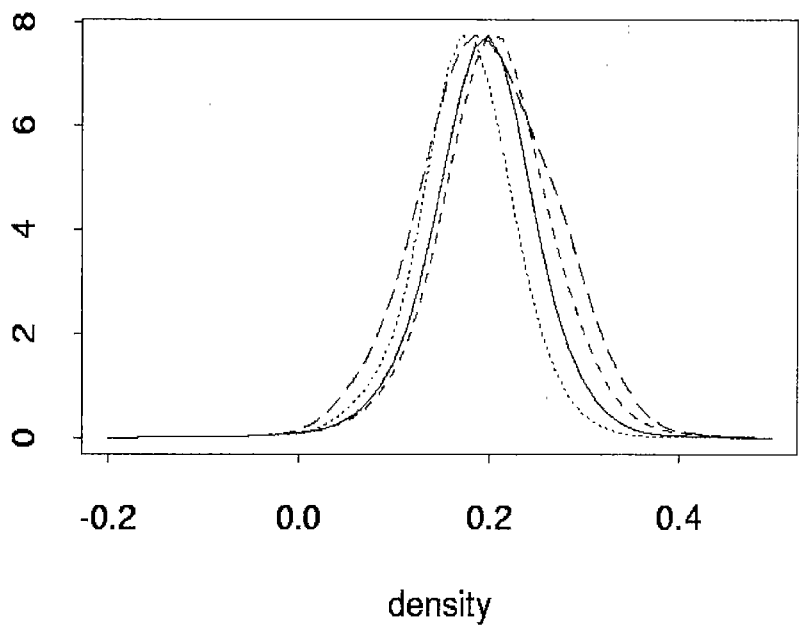

그림 4.1: μ의 추정된 사후분포

$M = 1$일 때의 평균 군집은 $\bar{k} = 1.5$가 된다. 또한 표 4.4는 M이 커짐에 따라 μ의 사후 추정치가 감소한다는 것을 알 수 있다. 만약 M이 무한대로 간다면 G는 기저분포인 G_0, 즉 정규분포로 가까이 가는 것을 의미한다. 표 4.4에 M을 10^4로 취하면 디리슈레 과정 사전분포는 근사적으로 정규분포로 가고 그에 따른 추정치들이 주어져 있다. 따라서 M이 클 때 우리의 결과는 Silliman의 결과와 유사해진다고 추측할 수 있다. 그러나 M이 랜덤이냐 아니냐에 상관없이 NaF가 SMFP보다 더 효과가 있다는 결론은 변하지 않는다.

5. 결론 및 앞으로의 과제

본 논문은 Silliman(1997)이 제안한 계층적 선택모형의 확장형태인 준모수적 계층적 선택모형을 소개하고 메타분석에 어떻게 적용되는지 보였다. 준모수적 계층적 선택모형은 메타분석이 본질적으로 가지고 있는 두 가지 문제점인 각 연구효과가 이질적이라는 것과 출판편의의 존재를 동시에 설명할 수 있다. 이 접근은 실제 적용에 매우 많이 쓰일 수 있는 모형이고 계산상의 어려움은 MCMC방법을 이용하여 해결하였다.

우리는 두 가지의 가중함수, $w(y) = |y|$와 $w(y) = |y|^2$와 정규 기저 사전분포하에서 전체적 평균 μ와 연구간 분산 σ_α^2을 추정하였다. 우리의 결과는 Johnson(1993)의 결과보다 어느정도 작은 값으로 나타나는 것으로 메타분석에서 자주 심각한 문제를 일으키는 출판 편의를 설명하였다. 그러나 NaF가 더 좋다는 결론은 변하지 않았다.

결론적으로 우리의 접근은 메타분석과 같은 문제를 해결하는 데 각 임의효과에 대한 가정을 약하게 하고 출판편의를 설명하는 것의 두 가지 잇점을 가지고 있다고 할 수 있다. 특히 우리의 예제처럼 표본의 수가 적을 때 완전한 모수적인 가정은 결론을 잘못 도출할 수 있는 문제점을 지니게 된다. 이것이 우리의 모형이 가지고 있는 가장 강한 잇점이라 할 수 있다.

앞으로의 과제는 다음과 같다. 첫째로 다양한 기저 사전분포를 고려해 볼 수 있다. 고려될 수 있는 기저분포로는 t분포나 로지스틱 분포같은 비정규 대칭분포와 치우친 정규분포(Azzalini와 Dalla-Valle, 1996; Chen, Dey와 Shao, 1999)와 같은 비대칭 분포를 고려해서 기저분포에 대한 로버스트성을 조사해 볼 수 있고 관심있는 모수인 μ와 σ_α^2에 대한 사전분포의 초월모수값의 변화에 따른 로버스트성도 조사해 볼 수 있다. 둘째로는 비 공액 사전분포 (non-conjugate prior)를 고려했을 때 일반적으로 발생하는 문제인 해석적으로 계산이 안 되는 q_0의 대체적인 계산(MacEachern과 Müller, 1998)을 고려해 보는 것과 마지막으로 계단 함수(step function)와 같은 가중함수의 다른 족(class)를 사용하여 자료분석을 하는 것도 의미가 있을 수 있다.

참고문헌

[1] Antoniak, C. E. (1974) Mixtures of Dirichlet Processes with Applications to Nonparametric Problems, *The Annals of Statistics*, 2, 1152-1174.

[2] Azzalini, A. and Dalla-Valle, A. (1996) The Multivariate Skew-Normal Distribution, *Biometrika*, 83, 715-726.

[3] Best, N. G., Cowles, M. K. and Vines, S. K. (1995) *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.3*, Cambridge, UK; MRC Biostatistics Unit.

[4] Chen, M. H., Dey, D. K. and Shao, Q. M. (1999) A New Skewed Link Model for Dichotomous Quantal Response Data, *Journal of the American Statistical Association*, 94, 1172-1186.

[5] Chib, S. and Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.

[6] Chung, Y. and Jeong, H. (2000) Bayesian Analysis of Grouped Random Effects Model in Meta-Analysis, *The Korean Journal of Applied Statistics*, 13, 81-96.

[7] DerSimonian, R. and Laird, N. (1986) Meta-Analysis in Clinical Trials, *Controlled Clinical Trials*, 7, 177-188.

[8] Escobar, M., D. and West, M. (1995) Bayesian Density Estimation and Inference using Mixtures, *Journal of the American Statistical Association*, 90, 577-588.

[9] Ferguson, T. S. (1973) A Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, 1, 209-230.

[10] 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.

[11] Geweke, J. (1992) Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments, in *Bayesian Statistics 4*, (J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith eds.), Oxford, UK; Oxford University Press, pp. 169-193.

[12] Hastings, W. K. (1970) Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 57, 97-109.

[13] Hedges. L. V. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*, New York: Academic Press.

[14] Johnson, M. F. (1993) Comparative Efficacy of Naf and SMFP Dentifrices in Caries Prevention : A Meta-Analysis Overview, *Journal of the European Organization for*

Caries Research (ORCA), 27, 328-336.

[15] Larose, D. and Dey, D. (1996) Weighted Distributions Viewed in the Context of Model Selection : a Bayesian Perspective, *Test*, 5, 227-246

[16] Larose, D. and Dey, D. (1997) Grouped Random Effects Models for Bayesian Meta-Analysis, *Statistics in Medicine*, 16, 1817-1829.

[17] MacEachern, S. N. and Müller, P. (1998) Estimating mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics*, 7, 223-238.

[18] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 21, 1087-1092.

[19] Morris, C. N. and Normand, S. L. (1992) Hierarchical Models for Combining Information and for Meta-Analysis, in *Bayesian Statistics 4*, (J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith eds.), Oxford, UK; Oxford University Press, pp. 321-344.

[20] Silliman, N. P. (1997) Hierarchical Selection Models with Applications in Meta-Analysis, *Journal of the American Statistical Association*, 92, 926-936.

[21] Tanner, M. A. and Wong, W. H. (1987) The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, 82, 528-550

[22] West, M. (1992) Hyperparameter Estimation in Dirichlet Process Mixture Models, *ISDS Discussion Paper #92-A03*, Duke University.

[23] West, M., Müller, P. and Escobar, M. D. (1994) Hierarchical Priors and Mixture Models, with Applications in Regression and Density Estimation, *Aspects of Uncertainty: A Tribute to D. V. Lindley* ed. A. F. M. Smith, and P. R. Freeman, London: John Wiley and Sons, pp. 363-386.

[2000년 6월 접수, 2001년 3월 채택]

A Bayesian Method to Semiparametric Hierarchical Selection Models

Younshik Chung[1] Junghoon Jang[2]

ABSTRACT

Meta-analysis refers to quantitative methods for combining results from independent studies in order to draw overall conclusions. Hierarchical models including selection models are introduced and shown to be useful in such Bayesian meta-analysis. Semiparametric hierarchical models are proposed using the Dirichlet process prior. These rich class of models combine the information of independent studies, allowing investigation of variability both between and within studies, and weight function. Here we investigate sensitivity of results to unobserved studies by considering a hierachical selection model with including unknown weight function and use Markov chain Monte Carlo methods to develop inference for the parameters of interest. Using Bayesian method, this model is used on a meta-analysis of twelve studies comparing the effectiveness of two different types of flouride, in preventing cavities. Clinical informative prior is assumed. Summaries and plots of model parameters are analyzed to address questions of interest.

Keywords: Bayesian meta-analysis; Clinical informative prior; Dirichlet process prior; Gibbs sampler; Hierarchical selection model; Metropolis algorithm; Mixture of Dirichlet process; Weight function.

1) Associate Professor, Department of Statistics, Pusan National University.
E-mail: yschung@hyowon.cc.pusan.ac.kr
2) Graduate, Department of Statistics, Pusan National University.