

로버스트 베이지안 메타분석[†]

최성미¹ · 김달호² · 신임희³ · 김호각⁴ · 김상경⁵

¹²경북대학교 통계학과 · ³대구가톨릭대학교 의학통계학과 ·
⁴대구가톨릭대학교 소화기내과 · ⁵대구가톨릭대학교 진단검사의학과

접수 2011년 4월 28일, 수정 2011년 5월 17일, 게재확정 2011년 5월 22일

요약

본 논문은 독립적으로 수행된 연구결과를 합쳐서 일반적인 결론을 도출하는 메타분석을 위한 로버스트 계층적 베이지안 모형을 고려했다. 사전정보가 정규분포를 따른다는 가정 대신 정규분포의 척도혼합을 사용하여 정규분포보다 더 두꺼운 꼬리를 가지는 사전분포를 사용한다. 나아가 개별 분석의 분산이 알려져 있지 않은 경우를 계층적 베이지안 모형에 포함하여 메타분석을 수행하고자 한다. 깁스 표집을 사용하여 추정값을 계산하고, 실제 자료를 사용하여 제안된 방법을 예시한다.

주요어: 계층적 베이지안, 깁스 표집, 로버스트, 메타분석, 정규분포, 척도혼합.

1. 서론

메타분석은 여러 독립적인 연구결과를 종합해서 전반적인 효과를 평가하거나 새로운 연구를 계획하는데 필요한 결론을 얻는 통계적 분석방법이다. 메타분석은 교육학, 심리학 뿐만 아니라 최근에는 의학 분야에서 매우 활발히 연구되고 있다. 예를 들면 동일한 치료효과를 평가하기 위해 여러 다른 센터에서 여러 임상연구들이 진행되었을 때 개별 연구는 표본의 크기 또는 대상자의 범위가 한정되어 있는 등 여러 가지 제약조건을 가지게 된다. 따라서 동일한 목적을 가지고 시행된 여러 연구 결과를 합리적이고 체계적으로 종합하는 것은 의미가 크다고 할 수 있다. 이를 위해 모형에 근거한 메타분석 방법이 많이 사용되는데, 이의 장점은 관심사인 연구들의 공통 평균의 추정이 모형에 의해 자동적으로 적절하게 가중되어 결합된다는 것이다. (DerSimonian과 Laird, 1986; Goodman, 1989; DuMouchel, 1990 참조).

여러 연구들이 존재할 때 이들 연구들로부터 공통효과 (common effect)의 추론에 관한 요약 정보를 도출하고자 한다. 메타분석에서 흔히 사용되는 랜덤효과모형은 계층적 구조를 가지므로 자연스럽게 베이지안 접근이 가능하게 된다. 베이지안 접근의 장점은 사전분포를 설정함으로써 개별 연구들이 가진 정보를 적절하게 연결시켜 공통효과를 보다 효과적으로 추론할 수 있다는 것이다.

본 논문에서는 메타분석에서 여러 연구들을 결합하고 통계적으로 요약하기 위하여 계층적 베이지안 모형을 고려했다. 베이지안 추론에서 사전분포를 사용함에 있어서 사전분포의 꼬리가 더 두꺼울수록 로버스트해지는 경향이 있다는 사실이 잘 알려져 있다. (Box와 Tiao, 1973; O'Hagan, 1979, 1989; West,

[†] 본 연구는 보건복지부 보건의료연구개발사업의 지원에 의하여 이루어진 것임 (과제고유번호 : A084177).

¹ (702-701) 대구시 북구 대학로 80번지, 경북대학교 통계학과, 박사과정.

² 교신저자: (702-701) 대구시 북구 대학로 80번지, 경북대학교 통계학과, 교수. E-mail: dalkim@knu.ac.kr

³ (705-718) 대구시 남구 대명4동 3056-6번지, 대구가톨릭대학교 의과대학 의학통계학과, 교수.

⁴ (705-718) 대구시 남구 대명4동 3056-6번지, 대구가톨릭대학교 의과대학 소화기내과, 교수.

⁵ (705-718) 대구시 남구 대명4동 3056-6번지, 대구가톨릭대학교 의과대학 진단검사의학과, 교수.

1987 참조). 정규분포의 척도혼합을 사전분포로 사용하면 정규분포의 꼬리보다 더 두꺼운 꼬리를 가지는 분포를 자연스럽게 만들 수 있다. 또한 정규분포의 척도혼합으로 만들어지는 분포의 집합에는 스튜던트 t 분포, 이중지수분포, 로지스틱 분포, 그리고 지수멱등족 (exponential power family) 등 상당히 많은 분포들이 포함된다.

한편, 개별 연구의 분산이 알려져 있지 않은 경우를 계층적 베이저안 모형에 포함하고자 한다. 이는 메타분석을 위한 보다 일반적인 방법을 고려하고자 하는 시도이다.

두꺼운 꼬리를 가지는 사전분포와 개별 연구의 분산이 알려져 있지 않은 경우를 모형에 포함하면 베이저안 추론을 위한 계산의 부담이 급증하게 된다. 이 문제를 극복하기 위해 최근에 잘 알려진 마코브 연쇄 몬테칼로 (Markov Chain Monte Carlo: MCMC) 방법을 사용하여 수치적 계산을 수행하고자 한다. (Carlin과 Polson, 1991; Gelfand와 Smith, 1991 참조).

계층적 베이저안 모형을 사용한 메타분석은 Goodman (1989)에 주어진 아스피린 자료를 이용하여 예시하고자 한다. 이 예제를 통해 메타분석에서 사전분포의 사용이 결과에 미치는 영향이 얼마나 다른지 보여준다.

2. 계층적 베이저안 모형

k 개의 연구에 대한 랜덤효과모형 $y_i = \theta_i + \epsilon_i$ ($i = 1, \dots, k$)를 고려하자. θ_i 는 알려지지 않은 모수로서 평균 구조를 의미하며 다음과 같이 표현된다.

$$\theta_i = \mu + \delta_i$$

여기서 δ_i 는 설명할 수 없는 연구간 변동을 의미하며, ϵ_i 는 평균이 0인 분포를 따르는 독립적인 랜덤오차이다. Morris와 Normand (1992)는 메타분석을 위한 계층적 베이저안 접근을 시도하였다. 여기서 관심사는 공통의 연구효과 μ 를 추정하는 것이다.

Morris와 Normand (1992)가 제안한 계층적 베이저안 접근은 계산의 편의를 위하여 연구내 분산인 V_i 를 알려진 값으로 가정하였으나, 만약 연구 내 분산 V_i 가 알려지지 않았다면 적절한 분포를 설정하여 계층적 베이저안 모형을 다음과 같이 표현할 수 있다.

$$(i) y_i \text{와 } \hat{V}_i \text{는 독립이며, } y_i | \theta_i \stackrel{\text{ind}}{\sim} N(\theta_i, V_i), \hat{V}_i \stackrel{\text{ind}}{\sim} V_i \frac{\chi_{n_i}^2}{n_i} \quad (i = 1, \dots, k);$$

$$(ii) \theta_i | V_1, \dots, V_k, \mu, \tau^2 \stackrel{\text{ind}}{\sim} N(\mu, \tau^2) \quad (i = 1, \dots, k);$$

$$(iii) \mu, \tau^2, V_1, \dots, V_k \text{은 서로 독립이며, } \mu \stackrel{\text{ind}}{\sim} \text{Uniform}(-\infty, \infty), \tau^2 \sim IG(\alpha, \beta), V_i \sim IG(\gamma, \delta).$$

여기서 $IG(\cdot, \cdot)$ 은 역감마 분포를 나타낸다. (만약 $f(x) \propto x^{-(\alpha+1)} \exp(-\beta x^{-1}) I_{(0, \infty)}(x)$ 이면 확률변수 X 는 $IG(\alpha, \beta)$ 분포를 따른다고 하며, 여기서 I 는 일반적으로 사용되는 지시함수를 일컫는다.) y_i 는 관찰된 연구효과를, θ_i 는 실제 연구효과, V_i 는 연구내 분산, \hat{V}_i 는 추정값의 표준오차, μ 는 공통의 연구효과, τ^2 는 연구간 분산을 의미한다.

위에서 자료와 모수들의 모형화를 위하여 정규사전분포를 가지는 계층적 베이저안 모형을 사용하였다. 그러나 이러한 제한된 분포의 사용은 제한적인 결과를 가져오며, 추론에 있어 부적합한 경우도 발생할 수 있다. 정규분포에 기초한 모형은 이상점에 로버스트 하지 않음은 잘 알려진 사실이다. 예컨대, 동떨어진 관찰값은 비록 실질적인 연관성이 적을지라도, 모형의 모든 모수를 추론하는 데 있어 상당히 강한 영향을 끼칠 수 있다.

이제 θ_i 의 사전분포로서 두꺼운 꼬리를 가지는 정규분포의 척도혼합을 사용하여 모형화하고자 한다. 로버스트화를 위한 베이저안 모형에 대한 배경은 West (1985)에 잘 소개되어 있다.

연구효과 θ_i 를 로버스트화 하기 위하여 $\theta_i | \mu, \tau^2, V_1, \dots, V_k \stackrel{ind}{\sim} N(\mu, \tau^2)$ 에서 정규분포 대신 보다 일반적인 분포로써 $\theta_i \sim \frac{1}{\tau} p\left(\frac{\theta_i - \mu}{\tau}\right)$ 를 사용하고자 한다. 여기서 $p(t) = \int_0^\infty \lambda^{1/2} \phi(t\lambda^{1/2})g(\lambda)d\lambda$ 이며, $\phi(t)$ 는 표준정규분포의 확률밀도함수를 나타낸다. 이때 $p(\cdot)$ 은 혼합분포 $g(\cdot)$ 을 가지는 정규분포의 척도혼합이다. 위의 식을 두 단계로 표현하면 다음과 같다.

- (i) $\theta_i | \mu, \tau^2, V_1, \dots, V_k \stackrel{ind}{\sim} N(\mu, \lambda_i \tau^2)$
- (ii) $\lambda_i \stackrel{ind}{\sim} g(\lambda_i)$

여기서 $\int_0^\infty g(t)dt = 1$ 이다. (Kim과 Park, 2000 참조).

정규성을 벗어나는 경우, 두꺼운 꼬리 사전분포를 얻기 위한 $g(\lambda)$ 의 함수적 형태의 목록은 다음과 같다. (West, 1987 참조).

(1) t 분포: 만약 $q/\lambda_i \sim \chi_q^2$ 이면 θ_i 는 자유도 q , 위치모수 μ , 척도모수 τ 를 가지는 스튜던트 t 분포를 따른다.

(2) 이중지수분포: 만약 λ_i 가 평균 1/2을 가지는 지수분포를 따르면 θ_i 는 위치모수 μ 와 척도모수 τ 를 가지는 이중지수분포를 따른다.

(3) 지수멱등축 분포: 만약 λ_i 가 지수 $\alpha/2$ 를 가지는 양의 안정분포 (positive stable distribution)를 따른다면 θ_i 는 위치모수 μ 와 척도모수 τ 를 가지는 지수멱등축 분포를 따른다.

(4) 로지스틱 분포: 만약 $1/\sqrt{\lambda_i}$ 가 점근적 콜모고로프 거리분포 (Kolmogorov distance distribution)를 따른다면 θ_i 는 위치모수 μ 와 척도모수 τ 를 가지는 로지스틱 분포를 따른다. (만약 $f(z) = 8z \sum_{j=1}^\infty (-1)^{j-1} j^2 \exp(-2j^2 z^2) I_{(0,\infty)}(z)$ 이면, 확률변수 Z 는 점근적 콜모고로프 거리 분포를 따른다고 한다.)

정규분포의 척도혼합을 적용한 로버스트 계층적 베이지안 모형은 다음과 같이 표현된다.

- (i) y_i 와 \hat{V}_i 는 독립이며, $y_i | \theta_i \stackrel{ind}{\sim} N(\theta_i, V_i)$, $\hat{V}_i \stackrel{ind}{\sim} V_i \frac{\chi_{n_i}^2}{n_i}$ ($i = 1, \dots, k$);
- (iia) $\theta_i | \mu, \tau^2, V_1, \dots, V_k \stackrel{ind}{\sim} N(\mu, \lambda_i \tau^2)$ ($i = 1, \dots, k$);
- (iib) $\lambda_i \stackrel{ind}{\sim} g(\lambda_i)$ ($i = 1, \dots, k$);
- (iv) μ 와 τ^2 와 V_1, \dots, V_k 은 서로 독립이며, $\mu \sim Uniform(-\infty, \infty)$, $\tau^2 \sim IG(\alpha, \beta)$, $V_i \sim IG(\gamma, \delta)$.

본 논문에서는 혼합분포 $g(\cdot)$ 를 다음 두 가지 사전분포로 사용하여 계산하였다.

- (i) $q/\lambda_i \sim \chi_q^2$
- (ii) $\lambda_i \sim \exp(1/2)$

메타분석에서 관심사는 모수 μ 와 τ^2 를 추정하는 데 있다. 이 경우 베이지안 추정량인 μ 와 τ^2 의 사후 평균과 분산을 구하는 것은 쉽지 않다. 왜냐하면 고차원의 적분문제가 발생하기 때문이다. 이러한 문제점을 해결하기 위해 마르코브 연쇄 몬테칼로 방법을 사용하여 사후분포를 생성하고 관련된 평균과 분산을 구할 수 있다. 구체적으로, 본 논문에서는 깃스 표집 (Gibbs sampling)을 사용하여 베이지안 추정량을 계산하고자 한다.

$\mathbf{y} = (y_1, y_2, \dots, y_k)$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, $\mathbf{V} = (V_1, V_2, \dots, V_k)$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_k)$ 라 두자. 깃스 표집을 사용하기 위해서는 먼저 \mathbf{y} 가 주어졌을 때 $\boldsymbol{\theta}$, \mathbf{V} , μ , $\boldsymbol{\lambda}$, τ^2 의 사후결합확률분포로부터 모든 모수에 대한 조건부확률분포를 구해야만 한다. 구체적으로 이들 조건부확률분포를 구하면 다음과 같다.

$$\begin{aligned}
\text{(i)} \quad & \theta_i | \mathbf{y}, \mathbf{V}, \mu, \tau^2, \boldsymbol{\lambda} \stackrel{\text{ind}}{\sim} N \left(\frac{y_i \tau^2 \lambda_i + V_i \mu}{\tau^2 \lambda_i + V_i}, \frac{V_i \tau^2 \lambda_i}{\tau^2 \lambda_i + V_i} \right) \quad (i = 1, \dots, k), \\
\text{(ii)} \quad & V_i | \mathbf{y}, \boldsymbol{\theta}, \mu, \tau^2, \boldsymbol{\lambda} \stackrel{\text{ind}}{\sim} IG \left(\frac{n_i + 1}{2} + \gamma, \frac{1}{2} \left\{ (y_i - \theta_i)^2 + n_i \widehat{V}_i + 2\delta \right\} \right) \quad (i = 1, \dots, k), \\
\text{(iii)} \quad & \mu | \mathbf{y}, \boldsymbol{\theta}, \mathbf{V}, \tau^2, \boldsymbol{\lambda} \sim N \left(\frac{\sum_{i=1}^k \theta_i \lambda_i^{-1} \tau^{-2}}{\sum_{i=1}^k \lambda_i^{-1} \tau^{-2}}, \frac{1}{\sum_{i=1}^k \lambda_i^{-1} \tau^{-2}} \right), \\
\text{(iv)} \quad & \lambda_i | \mathbf{y}, \boldsymbol{\theta}, \mathbf{V}, \mu, \tau^2 \stackrel{\text{ind}}{\sim} p(\lambda_i) \propto \lambda_i^{-1/2} \exp \left[-\frac{(\theta_i - \mu)^2}{2\tau^2 \lambda_i} \right] g(\lambda_i) \quad (i = 1, \dots, k), \\
\text{(v)} \quad & \tau^2 | \mathbf{y}, \boldsymbol{\theta}, \mathbf{V}, \mu, \boldsymbol{\lambda} \sim IG \left(\frac{k}{2} + \alpha, \frac{1}{2} \sum_{i=1}^k \frac{(\theta_i - \mu)^2}{\lambda_i} + \beta \right).
\end{aligned}$$

사전분포로 t 분포를 사용하는 경우를 고려하자. 로버스트 계층적 베이저안 모형의 (iii) 단계에서 혼합분포 $g(\cdot)$ 을 $q/\lambda_i \sim \chi_q^2$ 로 가정하자. 그러면 깃스 표집을 위한 조건부 확률분포의 (iv) 단계는 다음과 같이 표현된다.

$$\text{(iv)} \quad \lambda_i | \mathbf{y}, \boldsymbol{\theta}, \mathbf{V}, \mu, \tau^2 \stackrel{\text{ind}}{\sim} IG \left(\frac{q+1}{2}, \frac{q\tau^2 + (\theta_i - \mu)^2}{2\tau^2} \right) \quad (i = 1, \dots, k).$$

사전분포로 이중지수분포를 사용하는 경우, 혼합분포 $g(\cdot)$ 을 $\lambda_i \sim \exp(1/2)$ 로 가정하면 조건부 확률분포의 (iv) 단계는 다음과 같이 표현된다.

$$\text{(iv)} \quad \lambda_i | \mathbf{y}, \boldsymbol{\theta}, \mathbf{V}, \mu, \tau^2 \stackrel{\text{ind}}{\sim} p(\lambda_i) \propto \lambda_i^{-1/2} \exp \left\{ -2\lambda_i - \frac{(\theta_i - \mu)^2}{2\tau^2 \lambda_i} \right\} \quad (i = 1, \dots, k).$$

주어진 조건부 확률분포에 근거하여 깃스 표집을 이용한 계산을 수행할 때 깃스 표집의 수렴 여부를 쉽게 검증하기 위하여 길이가 $2d$ 인 L 개의 병렬체인을 고려한다. 각 체인에서 $\boldsymbol{\theta}, \mathbf{V}, \mu, \boldsymbol{\lambda}$ 와 τ^2 에 대해 $2d$ 개의 난수를 반복 생성한 후, 초기값의 영향을 제거하기 위하여 반복의 앞부분 d 개를 버리고 나머지 d 개를 사용하여 사후평균과 사후분산을 계산한다. 사후평균과 사후분산에 대한 계산의 정당성을 확인하기 위해 Gelman과 Rubin (1992)이 제안한 방법을 사용하여 깃스 표집의 수렴 여부를 검증한다. 이들이 제안한 절차에 따라 잠재적 척도축소인자 (potential scale reduction factor) \hat{R}_i 를 계산하였을 때, 만약 모든 관심모수에 대한 \hat{R}_i 가 1에 가까우면 이는 깃스 표집이 수렴한다는 것을 암시한다. 깃스 표집을 이용한 메타분석을 위한 로버스트 계층적 베이저안 추정량은 다음과 같이 추정된다.

$$\begin{aligned}
E(\mu | \mathbf{y}) &= E[E(\mu | \mathbf{V}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \tau^2, \mathbf{y})] \\
&\approx (Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} \left(\sum_{i=1}^k \lambda_i^{-(lm)} \tau^{-2(lm)} \right)^{-1} \left(\sum_{i=1}^k \theta_i^{(lm)} \lambda_i^{-(lm)} \tau^{-2(lm)} \right), \\
E(\tau^2 | \mathbf{y}) &= E[E(\tau^2 | \mathbf{V}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \mu, \mathbf{y})] \\
&\approx (Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} (k + 2\alpha - 2)^{-1} \left(\sum_{i=1}^k (\theta_i^{(lm)} - \mu^{(lm)})^2 \lambda_i^{-(lm)} + 2\beta \right).
\end{aligned}$$

한편 사후분산 $V(\mu | \mathbf{y})$ 와 $V(\tau^2 | \mathbf{y})$ 도 다음과 같이 추정된다.

$$\begin{aligned}
V(\mu | \mathbf{y}) &= E[V(\mu | \mathbf{V}, \lambda, \boldsymbol{\theta}, \tau^2, \mathbf{y})] + V[E(\mu | \mathbf{V}, \lambda, \boldsymbol{\theta}, \tau^2, \mathbf{y})] \\
&= E[V(\mu | \mathbf{V}, \lambda, \boldsymbol{\theta}, \tau^2, \mathbf{y})] + E[E(\mu | \mathbf{V}, \lambda, \boldsymbol{\theta}, \tau^2, \mathbf{y})^2] - (E[E(\mu | \mathbf{V}, \lambda, \boldsymbol{\theta}, \tau^2, \mathbf{y})])^2 \\
&\approx (Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} \left(\sum_{i=1}^k \lambda_i^{-(lm)} \tau^{-2(lm)} \right)^{-1} \\
&\quad + (Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} \left(\sum_{i=1}^k \lambda_i^{-(lm)} \tau^{-2(lm)} \right)^{-2} \left(\sum_{i=1}^k \theta_i^{(lm)} \lambda_i^{-(lm)} \tau^{-2(lm)} \right)^2 \\
&\quad + \left[(Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} \left(\sum_{i=1}^k \lambda_i^{-(lm)} \tau^{-2(lm)} \right)^{-1} \left(\sum_{i=1}^k \theta_i^{(lm)} \lambda_i^{-(lm)} \tau^{-2(lm)} \right) \right]^2,
\end{aligned}$$

$$\begin{aligned}
V(\tau^2 | \mathbf{y}) &= E[V(\tau^2 | \mathbf{V}, \lambda, \boldsymbol{\theta}, \mu, \mathbf{y})] + V[E(\tau^2 | \mathbf{V}, \lambda, \boldsymbol{\theta}, \mu, \mathbf{y})] \\
&= E[V(\tau^2 | \mathbf{V}, \lambda, \boldsymbol{\theta}, \mu, \mathbf{y})] + E[E(\tau^2 | \mathbf{V}, \lambda, \boldsymbol{\theta}, \mu, \mathbf{y})^2] - (E[E(\tau^2 | \mathbf{V}, \lambda, \boldsymbol{\theta}, \mu, \mathbf{y})])^2 \\
&\approx (Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} 2((k+2\alpha-2)^2(k+2\alpha-4))^{-1} \left(\sum_{i=1}^k (\theta_i^{(lm)} - \mu^{(lm)})^2 \lambda_i^{-(lm)} + 2\beta \right)^2 \\
&\quad + (Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} (k+2\alpha-2)^{-2} \left(\sum_{i=1}^k (\theta_i^{(lm)} - \mu^{(lm)})^2 \lambda_i^{-(lm)} + 2\beta \right)^2 \\
&\quad + \left[(Ld)^{-1} \sum_{l=1}^L \sum_{m=d+1}^{2d} (k+2\alpha-2)^{-1} \left(\sum_{i=1}^k (\theta_i^{(lm)} - \mu^{(lm)})^2 \lambda_i^{-(lm)} + 2\beta \right) \right]^2.
\end{aligned}$$

3. 수치분석

Goodman (1989)의 아스피린 자료를 사용하여 로버스트 베이저안 메타분석을 예시하고자 한다. 표 3.1은 미국과 유럽에의 6개 주요 다기관에서 시행된 연구에서의 심근경색 후 환자의 아스피린 대 위약에 대한 모든 원인으로부터의 사망률 차이 y_i 를 제공한다. 또한 평균 유효표본수 n_i (아스피린을 섭취한 수와 위약을 섭취한 수의 평균)와 각 연구에서의 추정된 표준오차 $\sqrt{\hat{V}_i} = \sqrt{\text{Var}(\hat{y}_i)}$ 도 나타나있다

표 3.1 아스피린 자료

연구 수준	UK-1	CDPA	GAMS	UK-2	PARIS	AMIS
n_i	619.50	764.50	313.00	841.00	608.00	2262.00
y_i (%)	2.77	2.50	1.84	2.56	2.31	-1.15
$\sqrt{\hat{V}_i}$ (%)	1.65	1.31	2.34	1.67	1.98	0.90

이 자료는 심근경색을 앓고 있는 환자들에 대한 아스피린의 보호 효과에 관한 6가지 연구의 결과를 나타낸다. 각 연구에서의 효과는 위약 그룹과 아스피린 그룹 사이의 사망률의 차이로 측정되었다. 6개의 연구 가운데 5개에 대해서는 일반적인 수준에서 유의하지는 않지만 양의 부호로 나타났고, 대조적으로 6번째의 가장 큰 연구에서는 반대의 부호를 가진다.

아스피린의 효과의 평가에 대한 분석을 위한 계층적 모형에서 정규 사전분포 가정 대신 효과 θ_i 의 긴 꼬리 사전분포를 사용하고자 한다. 우리의 관심은 이러한 연구에 대해 공통효과 μ 를 추정하는 데 있다.

정규분포의 척도혼합을 사용한 두꺼운 꼬리 사전분포에 기초하여 모수의 로버스트 베이지안 추정을 유도하는 데 있어, 우리는 깃스 표집의 수집 여부를 쉽게 검증하기 위하여 10개의 독립적인 병렬체인을 가지는 깃스 표집을 고려하였다. 또한 각 체인에서 $2d = 10,000$ 개의 반복을 사용하여 θ , V , μ , λ 와 τ^2 의 난수를 각각 10,000개 생성한 후, 초기값의 영향을 제거하기 위해 반복의 앞부분 5,000개를 버리고 나머지 5,000개를 사용하여 사후평균과 사후분산을 계산하였다.

표 3.2는 정규분포, 이중지수분포, 그리고 자유도 1, 3, 5, 10의 t-분포를 사전분포로 가지는 경우에 대한 μ 와 τ^2 의 베이지안 추정값으로 사후평균과 사후표준편차를 제공한다. 분산의 역수에 대해 흩어진 형태의 감마 사전분포 (diffuse gamma prior)를 설정하기 위해 α 와 β , γ , δ 의 값을 모두 0.01로 사용하였다.

표 3.2를 살펴보면 사전분포의 꼬리가 두꺼워질수록 베이지안 추정값은 표본평균 $\bar{y} = 1.805$ 에 가까워짐을 알 수 있다. 그리고 사전분포의 꼬리가 얇아질수록 베이지안 추정값이 정규사전분포 하에서의 베이지안 추정값에 가까워짐을 알 수 있다. 이중지수사전분포 하에서의 베이지안 추정값은 t_3 과 t_5 사전분포 하에서의 추정값 사이에 있음을 알 수 있다. 이는 이중지수분포는 $t_\nu, \nu = 1, 2, 3$ 보다는 꼬리가 얇고, $t_\nu, \nu = 5, 6, \dots$ 보다는 꼬리가 두껍기 때문이다. 이 때 베이지안 추정값의 사후표준편차 (혹은 표준오차)가 꼬리가 두꺼울수록 커지는 것은 로버스트 추정을 위해 흩어진 형태의 사전정보를 사용했기 때문인 것으로 사료된다.

표 3.2 아스피린 자료에 대한 μ 와 τ^2 의 베이지안 추정값

	μ		τ^2	
	추정값	표준오차	추정값	표준오차
t(1)	1.757812	0.9983311	0.8464913	2.843708
t(3)	1.523490	0.9665383	1.7259590	4.394074
t(5)	1.468097	0.9605904	2.0895320	4.744727
t(10)	1.392884	0.9558240	2.3315550	5.664950
정규분포	1.371330	0.9381180	2.6396200	6.823126
이중지수분포	1.491677	0.9787772	1.9585110	5.773235

Gelman과 Rubin (1992)은 깃스 표집을 사용하여 구한 사후평균과 사후분산에 대한 계산의 정당성을 확인하기 위하여 깃스 표집기 (Gibbs sampler)의 수렴성을 조사하는 절차를 제안하였다. 만약 제안된 절차를 통해 계산된 잠재적 척도축소인자 \widehat{R}_i 이 모두 1에 가까우면 깃스 표집이 수렴한다고 판단할 수 있다.

사후예측 P 값은 모형의 적합성을 검증하는데 유용하다. 사후예측 P 값을 이용하여 모형적합을 확인하는 원리는 주어진 모형에 근거하여 새롭게 생성된 자료가 관측된 자료와 비슷한 결과가 나타나는지를 확인하는 것이다. 적절한 불일치측도의 모의실험 값이 사후예측분포로부터 생성되며 이를 관측된 데이터에 대한 불일치측도 값과 비교한다. (Sinha와 Dey, 1997 참조). 사후예측 P 값을 계산했을 때 0 이나 1에 가까운 극단적인 값을 가지면 모형의 적합결여를 나타낸다고 판단하는 반면, 모형이 자료를 잘 적합하면 이 P 값은 0.5에 가깝게 나타난다.

본 논문의 경우, 두 가지 경우 모두 모든 θ_i 에 대해 잠재적 척도축소인자 $\widehat{R}_i \approx 1$ 이며, 사후예측 P 값 또한 $P \approx 0.5$ 로 나타나므로 깃스 표집이 잘 수렴하며 모형도 적합하다고 판단되었다.

참고문헌

- Box, G. E. P and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*, John Wiley & Sons, New York.

- Carlin, B. P. and Polson, N. G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of Statistics*, **19**, 399-405.
- DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trial*, **7**, 177-188.
- DuMouchel, W. H. (1990). Bayesian meta-analysis. In *Statistical Methodology in the Pharmaceutical Sciences*, edited by D. A. Berry, Marcel Dekker, New York, 509-529.
- Goodman, S. N. (1989). Meta-analysis and evidence. *Controlled Clinical Trials*, **10**, 188-204.
- Gelfand, A. E. and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations. *Communication in Statistics - Theory and Methods*, **2**, 1747-1766.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-511.
- Kim, D. H. and Park, G. J. (2000). Robust Bayesian models for meta-analysis. *Journal of the Korean Data & Information Science Society*, **11**, 313-318.
- Morris, C. N. and Normand, S. L. (1992). Hierarchical models for combining information and for meta-analyses. In *Bayesian Statistics 4*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, 321-344.
- O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *The Journal of Royal Statistical Society, Ser. B*, **4**, 358-367.
- O'Hagan, A. (1989). Modeling with heavy tails. In *Bayesian Statistics 3*, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindely and A. F. M. Smith, Oxford University Press, 349-359.
- Sinha, D. and Dey, D. (1997). Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, **92**, 1195-1212.
- West, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2*, edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindely and A. F. M. Smith, North Holland, 531-557.
- West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, **7**, 646-648.

Robust Bayesian meta analysis[†]

Seong Mi Choi¹ · Dal Ho Kim² · Im Hee Shin³ · Ho Gak Kim⁴ · Sang-Gyung Kim⁵

¹²Department of Statistics, Kyungpook National University

³Department of Medical Statistics, Catholic University of Daegu

⁴Department of Internal Medicine, Catholic University of Daegu

⁵Department of Laboratory Medicine, Catholic University of Daegu

Received 28 April 2011, revised 17 May 2011, accepted 22 May 2011

Abstract

This article addresses robust Bayesian modeling for meta analysis which derives general conclusion by combining independently performed individual studies. Specifically, we propose hierarchical Bayesian models with unknown variances for meta analysis under priors which are scale mixtures of normal, and thus have tail heavier than that of the normal. For the numerical analysis, we use the Gibbs sampler for calculating Bayesian estimators and illustrate the proposed methods using actual data.

Keywords: Gibbs sampling, hierarchical Bayesian, meta-analysis, normal, robust, scale mixture.

[†] This study was supported by a grant of the Korea Healthcare Technology R&D Project, Ministry for Health & Welfare, Republic of Korea (Project No.: A084177).

¹ Doctor of philosophy student, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.

² Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: dalkim@knu.ac.kr

³ Professor, Department of Medical Statistics, School of Medicine, Catholic University of Daegu, Daegu 705-718, Korea.

⁴ Professor, Division of Gastroenterology, Department of Internal Medicine, School of Medicine, Catholic University of Daegu, Daegu 705-718, Korea.

⁵ Professor, Department of Laboratory Medicine, School of Medicine, Catholic University of Daegu, Daegu 705-718, Korea.