응용통계연구 제 18권 1호, 2005년, pp. 1–13

잠재그룹 포아송 모형을 이용한 전립선암 환자의 베이지안 그룹화 *

오만숙[1)]

요 약

최근 많은 연구자와 실무자들이 모집단에 내재해 있는 여러 다른 그룹(class, segment)간의 이질성을 밝혀내고 객체들을 그룹별로 세분화하는 방법 중 하나로 잠재그룹 모델(Latent class model)을 고려하고 있다. 이 논문에서는 2000년도에 국립 암 센터에 접수된 한국 내 연령별 전립선암 사망자수 자료를 기반으로, 잠재그룹 포아송 모형을 이용하여 전립선암 환자의 연령에 따른 그룹화를 시도한다. 최우추정법 등 고전적 추론방법의 한계를 극복하기 위하여 Markov Chain Monte Carlo (MCMC) 방법을 도구로 한 베이지안 추정방법을 제안한다. 제안된 베이지안 방법의 장점은 용이한 모수추정과 추정오차의 제공, 그리고 각 객체의 소속그룹의 판정과 이에 따르는 오차, 즉, 객체의 각 군집에 속할 확률, 도 구할 수 있다는 것이다. 또한 주어진 자료들에 대해 가장 적합한 그룹의 수를 결정하는 방법을 제시하여 그룹의 수나 세분화의 근거를 사전에 제공하지 않아도 자료가 주는 정보로부터 이들을 자동으로 결정하는 방법을 제시한다.

주요용어: 잠재그룹 모형, 합성모형, 마코프 체인 몬테칼로, 최적 모형 선택

1. 서론

최근에 모집단에 존재하는 서로 이질적인 그룹(class, segment)을 밝혀내기 위한 방법으로 잠재그룹 모형 (Latent Class model : LC model)이 자주 이용되고 있다 (DeSarbo and Cron, 1988; Kamakura and Russel, 1989; DeSoete and DeSarbo, 1991; Jedidi, Ramaswamy and DeSarbo, 1993; DeSarbo and Choi; 1999). 이는 다른 객체 수준별 분석 모델에 비하여 LC 모델이 가지는 장점들 때문으로, 특징적인 장점들로는 다음과 같은 것들이 있다. 첫째, LC 모델은 그룹화와 그룹-수준별 모수 추정(segment-level parameter estimation)을 동시에 구하도록 고안되었다. 각 객체가 어느 그룹에 속하는지를 알려줄 뿐만 아니라 각 그룹의 분포에 대한 개별 모형의 모수도 추정하여 주므로 각 그룹의 특성을 잘 파악할 수 있는 것이다. 둘째; LC 모델은 미리 그룹화의 근거를 제공하지 않고 데이터가 가진 정보를 이용하여 스스로 그룹화를 수행하는 데이터-의존성(data-driven)을 지닌 자동 그룹화(unsupervised clustering)를 수행한다. 이는 실제적인 경우에 어느 근거로 모집단이 세분화되는지 알지 못

* 본 연구는 한국과학재단 목적기초연구(R06-2002-012-01002-0)지원으로 수행되었음

1) (120-750) 서울시 서대문구 대현동 21 이화여자대학교 통계학과 교수
E-mail: msoh@mm.ewha.ac.kr.

하는 경우가 많기 때문에 매우 유용한 특징이다. 끝으로, LC 모형은 각 객체의 그룹에 대한 소속확률(membership probability)을 제공한다. 대부분의 그룹화 방법은 각 객체의 소속확률을 제공하지 않는데, 각 객체의 소속확률을 모르는 경우 문제점은 소속확률들이 거의 비슷할 때, 즉, 객체의 소속이 분명치 않을 때, 더욱 두드러진다. 예를 들어, 어떤 객체가 그룹 1에 속할 확률이 0.45이고 그룹 2에 있을 확률이 0.55하고 가정하자. 그룹 2에 대한 소속확률이 더 크므로 주어진 객체는 그룹 2에 속한다고 결론을 내릴 것이다. 그러나 이 경우에 주어진 객체는 그룹 1과 그룹 2에 대한 소속확률이 거의 차이가 없기 때문에 그룹 2에 속한다고 분류하는 것이 의미가 없다. 소속확률을 모르는 경우 위와 같은 경우를 감지할 수 없고, 따라서 예를 들면 어느 객체가 확률 1로 그룹 2에 속하는지 확률 0.51로 그룹 2에 속하는지 구분할 수가 없는 것이다. 반면에 LC 모델은 각 객체의 소속확률을 제공함으로써 객체들의 분류 혹은 그룹화에 따르는 불확실성의 측도를 제공한다는 장점이 있다.

본 논문에서는 2000년도에 국립 암 센터에 접수된 한국 내 연령별 전립선암 사망자수 자료를 기반으로, 잠재그룹 포아송 모형을 이용하여 전립선암 환자의 연령에 따른 그룹화를 제안한다. 이 자료는 국립 암 센터의 사망 원인 통계 연보로부터의 얻은 것으로, 연령을 1-84세까지 5년 단위로 나누고 85세 이상을 하나로 하여 총 18개의 각 연령 그룹에 대하여 2000년도에 전립선암으로 인한 사망자 수를 기록한 것이다. 참고로 이 자료는 호적법 및 통계법에 따라 국민이 신고한 사망 신고서의 사망 원인을 기초로 수집된 자료이다.

표 5.1에 나타난 자료를 보면 편의상 연령을 5년 단위로 나누어 조사하였기 때문에 총 18개의 연령 그룹이 존재하나, 자세히 들여다보면 사망자 수에 있어서 거의 차이가 없는 연령 그룹이 존재함을 짐작할 수 있다. 이는 전립선 암 사망자수를 중심으로 연령을 그룹화할 때 18개 보다 작은 수의 그룹으로 묶는 것이 가능함을 알 수 있게 한다. 이들 그룹들이 각 그룹 내에서는 동질적인(homogeneous) 객체들로 구성되어 있으나 서로 다른 그룹들끼리는 서로 이질적인(heterogeneous) 성질을 갖게 될 것이다. 따라서 잠재그룹 모형에서 같은 그룹으로 분류된 연령들은 별도로 취급할 필요 없이 같이 묶어서 조사 혹은 병의 관리를 하면 될 것이다.

이후의 본 논문은 다음과 같이 구성된다. 2장은 잠재그룹 포아송 모형을 소개하고 잠재그룹 모형의 추정에 있어서 고전적 최우 추정의 문제점과 그 해결책으로서 베이지안 방법을 제안한다. 3장에서는 잠재그룹 모형의 모수에 대한 사전분포와, 베이지안 추정에 필요한 MCMC 기법의 사용에 필요한 각 모수의 조건부 사후분포를 유도한다. 4장에서는 적절한 그룹의 수를 결정하기 위한 베이지안 방법을 기술하고 5장에서는 전립선암 자료에 대한 베이지안 그룹화 분석을 제시한다. 마지막 장은 요약과 결론에 할애하였다.

2. 잠재그룹 포아송 모형

관측치 Y_i는 양의 정수를 가지며 독립적으로 다음과 같은 잠재그룹 포아송 모형을 따른다고 가정한다.

$$Y_i \sim \sum_{k=1}^{K} \epsilon_k \cdot P(\lambda_k) \ , \ i = 1, \ldots, n.$$

여기에서 ϵ_k는 k번째 그룹이 전체에서 차지하는 비율이고 $P(\lambda_k)$는 평균 λ_k를 갖는 포아송 분포를 나타내며, K는 총 그룹의 수이다. 변수 J_i를 i번째 관측치의 membership indicator로, 즉, Y_i가 그룹 k에 속한다면 $J_i = k$라고 정의한다. J_i를 사용하면 변수 $Y = (Y_1 \dots Y_n)$의 결합 확률 밀도 함수(joint probability density function)는

$$f(y) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\epsilon_k \frac{e^{-\lambda_k} \lambda_k^{y_i}}{y_i!} \right)^{I(J_i=k)} \tag{2.1}$$

으로 $I(\cdot)$는 지시함수이다.

이처럼 각 관측치에 대하여 합성모형을 가정하는 LC 모형은 1장에서 언급한 중요한 장점들이 있는 반면 실제 추정에 있어서 다음과 같은 문제점을 지닌다. 첫째, 모수의 추정이 용이하지 않다는 것이다. 모수의 수가 많고 우도함수가 단순하지 않기 때문에 특히 최우추정법을 이용한 모수 추정은 심각한 복잡성을 지닌다. 또 다른 문제점은 적절한 그룹(class)의 수를 어떻게 정하느냐 하는 문제이다. 이는 통계적 모형선택의 문제로 볼 수 있는데 주로 사용되는 최우추정치(MLE)에 기반한 우도비 검정(Goodman, 1974; Formann, 1985)은 이 경우에 적절하지 않다. 왜냐하면 많은 수의 그룹(class)을 갖는 LC 모형이 작은 수의 그룹을 갖는 LC 모형을 내포하지 않아 우도비에서 필요로 하는 정규 조건이 LC 모델에서 만족되지 않기 때문이다 (Heinen, 1993; Hoijtink, 1998; DeSarbo and Choi, 1999).

본 논문에서는 잠재그룹 모델에 대한 이러한 문제점에 대한 해결방안으로 베이지안 방법을 고려한다. LC 모형의 모수 추정을 위하여 MCMC 방법의 하나인 깁스 샘플링 알고리즘(Gibbs sampling algorithm; Gelfand and Smith, 1990)을 사용한다. 적절한 사전분포를 사용하면 3장에서와 같이 각 모수의 조건부 사후분포를 매우 편리한 형태로 얻을 수 있고 따라서 깁스 샘플링 알고리즘을 사용하여 모수들의 사후표본을 쉽게 얻을 수 있다. 이 사후표본으로부터 모수의 추정치와 추정오차를 계산할 수 있고 나아가 각 개체의 그룹들에 대한 소속확률을 얻을 수 있다. 끝으로 적절한 그룹의 수를 결정하는 문제를 통계적 모형선택의 문제로 바꾸어 서로 다른 개수의 계층을 갖는 LC 모형들에 대한 사후확률을 계산하고 이를 근거로 적절한 모형, 즉, 적절한 그룹의 수, 을 결정하는 베이지안 접근법을 사용한다. 모형의 사후확률 계산은 수리적으로 불가능하므로 이를 Oh(1999) 가 제안한 사후밀도 함수의 추정을 통하여 수치적으로 계산하는 방법을 사용한다.

3. 사전분포와 조건부 사후분포

3.1. 사전 분포

모수들에 대한 사전 분포를 위해서, 우리는 $\epsilon = (\epsilon_1 \dots \epsilon_K)$과 $\lambda = (\lambda_1 \dots \lambda_K)$가 서로 독립이라고 가정한다. 또한 $\{\lambda_k, k = 1, \dots, K\}$ 들 사이에서도 독립이라고 가정한다. 부적합 사전 밀도 함수(improper priors)는 베이지안 모형 선택에서 문제를 야기할 수 있으므로, 우리는 모수들에 대하여 다음과 같은 공액 사전 분포를 가정한다.

$$\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_K) \sim D(1, \dots, 1) \tag{3.1}$$

$$\lambda_k \sim iid \quad Gamma(\alpha, \beta)\ , \quad k = 1, \ldots, K \tag{3.2}$$

여기에서 ϵ 에 대한 모호 사전 분포(vague prior)는 모수 $(1, \ldots, 1)$을 가진 드리슈레 분포(Dirichlet distribution) 이고, λ에 대한 공액 사전 분포는 모수 (α, β)를 가진 감마 분포이다. 또한 이 모형에서 λ에 대한 모호성을 주기 위해서 $\alpha = 1, \beta = 0.01$이라고 가정하여 사전평균 100, 사전표준편차 100을 갖도록 한다.

3.2. 조건부 사후분포

우리는 식(2.1)에 Y의 확률밀도함수를 나타내었다. 그리고 식(3.1)과 식(3.2)에서 기술된 사전분포로부터 결합 사전분포를 구하면

$$\begin{aligned} \pi(\lambda, \epsilon) &= \pi(\lambda_1)\pi(\lambda_2)\ldots\pi(\lambda_K)\pi(\epsilon) \\ &= \prod_{k=1}^{K}\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda_k^{\alpha-1}e^{-\beta\lambda_k}\right) \times \frac{\Gamma(K)}{\prod_{k=1}^{K}\Gamma(1)}\prod_{k=1}^{K}\epsilon_k^{1-1} \\ &= \Gamma(K)\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^K \prod_{k=1}^{K}(\lambda_k^{\alpha-1}e^{-\beta\lambda_k}) \end{aligned} \tag{3.3}$$

이 된다. 다음, 식(2.1)의 우도함수와 식(3.3)의 사전밀도함수로부터 사후밀도함수(posterior density function)를 구하면,

$$\begin{aligned} \pi(\lambda, \epsilon|y) &\propto f(y|\lambda, \epsilon, K)f(K)\pi(\lambda, \epsilon) \\ &\propto \prod_{i=1}^{n}\prod_{k=1}^{K}(\epsilon_k e^{-\lambda_k(1+\beta)}\lambda_k^{y_i+\alpha-1})^{I(J_i=k)} \end{aligned} \tag{3.4}$$

임을 알 수 있다.

식(3.4)로부터, 우리는 쉽게 다음과 같이 각 모수에 대한 조건부 사후 분포를 얻어낼 수 있다.

$$\pi(\lambda_k|\epsilon, y, \{J_i\}) \sim Gamma(\sum_{i=1}^{n} y_i I(J_i = k) + \alpha, n_k + \beta) \tag{3.5}$$

$$\pi(\epsilon|\lambda, y, \{J_i\}) \sim D(n_1 + 1, \ldots, n_K + 1) \tag{3.6}$$

$$P(J_i = k|\lambda, \epsilon, y) = \frac{\epsilon_k f_k(y_i)}{\sum_{k=1}^{K}\epsilon_k f_k(y_i)} \tag{3.7}$$

이며 여기에서 $n_k = \sum_{i=1}^{n} I(J_i = k)$, 즉, 그룹 k에 속하는 자료의 수를 말하고, $f_k(y_i) = \frac{\lambda_k^{y_i}e^{-\lambda_k}}{y_i!}$이다.

각각의 모수들에 대한 조건부 사후 분포가 주어졌으므로 깁스 샘플링 (Gibbs sampling) 알고리즘을 이용하여 모수의 사후표본을 얻을 수 있고 이 사후표본으로부터 모수의 추정치와 추정오차를 계산할 수 있다. 모수의 추정치가 얻어지면 식 (3.7) 에 모수의 추정치를 대입하여 각 개체의 그룹에 대한 소속확률을 구할 수 있다.

4. 잠재그룹 수의 결정

이 장에서, 우리는 잠재그룹 모형에서 적절한 그룹의 수, K, 를 결정하는 방법을 제안하고자 한다. 적절한 그룹의 수를 정하는 문제를 베이지안 모형 선택의 관점에서 보면, 자료 y가 주어졌을 때 K의 사후확률 $\pi(K|y)$를 최대화시키는 K 가 바로 가장 적절한 그룹의 수가 될 것이다. 사후확률 $\pi(K|y)$를 직접적으로 계산하기 어려우므로 다음 식

$$\pi(K|y) \propto \pi(y|K)\pi(K)$$

을 고려한다. K의 가능한 값으로 $1 \leq K \leq \max K$ 을 가정하고 사전확률 $\pi(K)$에 대하여 $\pi(K) = \frac{1}{\max K}$ 로 균등확률을 가정하면,

$$\begin{aligned} \pi(K|y) &\propto \pi(y|K)\pi(K) \\ &\propto \pi(y|K) \end{aligned} \tag{4.1}$$

가 되고 따라서 K 가 주어졌을 때 자료 y의 주변우도함수(marginal likelihood function) $\pi(y|K)$를 최대화 시키는 K값을 찾으면 될 것이다. 다음, $\pi(y|K)$를 구하기 위해서 다음과 같은 식을 사용한다. K가 주어졌을 때 LC 모형의 모수를 θ 로, 즉 , $\theta = (\lambda, \epsilon)$로 놓는다. 베이즈 정리로부터

$$\pi(\theta|y, K) = \frac{f(y|\theta, K)\pi(\theta|K)}{\pi(y|K)}$$

이고 따라서

$$\pi(y|K) = \frac{f(y|\theta, K)\pi(\theta|K)}{\pi(\theta|y, K)} \tag{4.2}$$

를 얻을 수 있다. 식(4.2)에서 θ 값은 임의의 값이 들어갈 수 있으므로 깁스 샘플링에서 얻은 모수들의 추정값 $\theta^* = (\lambda^*, \epsilon^*)$를 대입하면

$$\pi(y|K) = \frac{f(y|\theta^*, K)\pi(\theta^*|K)}{\pi(\theta^*|y, K)}$$

이다. 따라서

$$log\pi(y|K) = logf(y|\theta^*, K) + log\pi(\theta^*|K) - log\pi(\theta^*|y, K) \tag{4.3}$$

을 얻는다. 식(2.1)과 식(3.3)로부터 $f(y|\theta^*, K)$와 $\pi(\theta^*|K)$를 다음과 같이 쉽게 구할 수 있다.

$$f(y|\lambda^*, \epsilon^*, K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \left(\epsilon_k^* \frac{e^{-\lambda_k^*} \lambda_k^{*y_i}}{y_i!} \right) \tag{4.4}$$

$$\pi(\lambda^*, \epsilon^*|K) = \Gamma(K) \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \right)^K \prod_{k=1}^{K} (\lambda_k^{*\alpha-1} e^{-\beta\lambda_k^*}) \tag{4.5}$$

이제, $\pi(\theta^*|y,K)$를 구하면 된다. Oh(1999)는 깁스 샘플링 알고리즘으로부터 얻은 사후 표본을 이용하여 사후 밀도 함수 $\pi(\theta^*|y,K)$를 구하는 방법을 다음과 같이 제안하였다.

$$\begin{aligned}\pi(\theta^*|y,K) &= \pi(\lambda^*,\epsilon^*|y,K) \\ &= E[\pi(\lambda^*|\epsilon^*,y,\{J_i\},K)\pi(\epsilon^*|\lambda,y,\{J_i\},K)] \\ &= \frac{1}{N}\sum_{j=1}^{N}[\pi(\lambda^*|\epsilon^*,y,\{J_i\}_j,K)\pi(\epsilon^*|\lambda_j,y,\{J_i\}_j,K)]\end{aligned} \tag{4.6}$$

로, 여기에서 λ_j는 $\lambda=(\lambda_1,\ldots,\lambda_K)$의 j번째 사후표본, $\{J_i\}_j$ 는 $\{J_i\}$의 j번째 사후표본을 나타낸다. 식(3.5), 식(3.6)에서 모수들에 대한 조건부 분포가

$$\pi(\lambda_k^*|\epsilon^*,y,\{J_i\},K) \sim Gamma(\sum_{i=1}^{n} y_i I(J_i=k)+\alpha, n_k+\beta)$$

$$\pi(\epsilon^*|\lambda,y,\{J_i\},K) \sim D(n_1+1,\ldots,n_K+1)$$

로 주어지므로 식(4.6)을 이용하여 손쉽게 $\pi(\theta^*|y,K)$를 추정할 수 있다. 이 추정치를 식(4.3) 에 대입하여 $log\pi(y|K)$를 최대화 시키는 K값을 얻을 수 있다. 이 K값이 잠재그룹 포아송 모형에서 적절한 그룹의 수가 될 것이다.

5. 전립선암 환자의 그룹화

5.1. 자료 탐색

이 자료는 국립 암 센터의 사망 원인 통계 연보로부터 얻은 것으로, 연령을 1-84세까지 5년 단위로 나누고 85세 이상을 하나로 하여 총 18개의 각 연령 그룹에 대하여 2000년도에 전립선암으로 인한 사망자 수를 기록한 것이다. 참고로 이 자료는 호적법 및 통계법에 따라 국민이 신고한 사망 신고서의 사망 원인을 기초로 한 것이다. 표 5.1에 2000년도 전체 전립선 암 사망자수 548명의 연령별 분포가 요약되어 있다. 그림 5.1은 이 분포를 막대그래프로 표시한 것으로 가로축은 연령그룹, 세로축은 사망자수를 나타낸다. 또한 그림 5.2는 전립선 암 사망자수에 대한 상대 그룹수(그룹수/전체 그룹수)의 히스토그램과 잠재그룹 포아송 모형을 이용하여 추정한 밀도함수를 나타낸다.

이 자료에서 각 연령그룹 마다 관측된 전립선 암 사망자 수는 잠재그룹 포아송 모형을 따른다고 가정한다. 잠재그룹의 수에 따라 연령별 사망자수의 모집단 분포가 달라질 것이다. 예를 들면, 잠재그룹을 1로 하면 연령별로 조사된 18개의 사망자수가 하나의 포아송 분포로부터 얻은 18개의 랜덤 관측치로 간주되는 것으로 전체 모집단이 동질적임을 의미한다, 즉, 연령에 따른 영향이 없다고 볼 수 있다. 반면, 만약 잠재그룹의 수를 18로 한다면 각 연령마다 서로 다른 평균을 갖는 포아송 분포를 가정하는 것으로 완전한 이질성을 의미한다, 즉, 주어진 연령 그룹들이 공통점이 없이 개별적인 분포를 갖는 것이다. 자료를 살펴보

표 5.1: 2000년도 연령별 전립선 암 사망자수

obs.	연령별	사망자수
1	1-4세	0
2	5-9세	0
3	10-14세	0
4	15-19세	0
5	20-24세	0
6	25-29세	0
7	30-34세	1
8	35-39세	2
9	40-44세	0
10	45-49세	5
11	50-54세	10
12	55-59세	20
13	60-64세	36
14	65-69세	78
15	70-74세	100
16	75-79세	113
17	80-84세	104
18	85세 이상	79

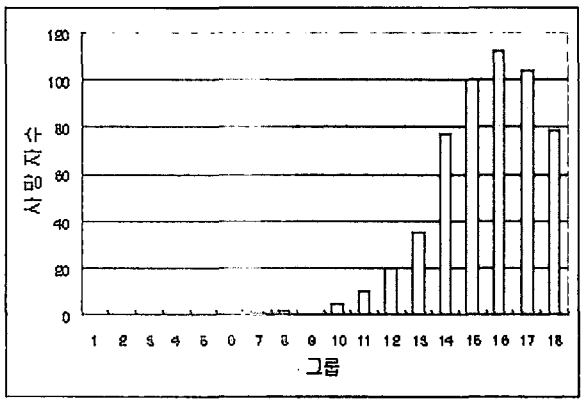

그림 5.1: 2000년도 연령 그룹별 전립선암 사망자수에 대한 히스토그램

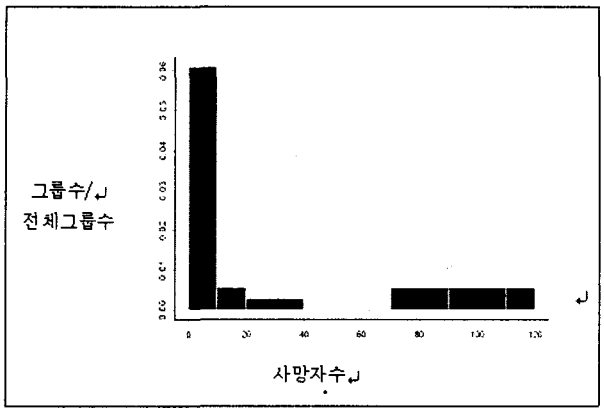

그림 5.2: 전립선 암 사망자수에 대한 상대 그룹수(그룹수/전체 그룹수)의 히스토그램과 밀도함수

면 18개의 연령 그룹이 편의상 5년 단위로 나뉘어져 있을 뿐 사망자수를 고려한 그룹핑은 아닌 것으로 보인다. 사망자수를 고려한 그룹들 간의 유사성을 살펴보면 몇 개의 연령그룹들, 특히 젊은 층 그룹들, 은 비슷한 수의 사망자수를 가지고 있어 유사성을 지닌 반면 젊은 층과 노년층 사이에는 뚜렷한 차이, 즉, 이질성을 가짐을 알 수 있다. 따라서 적절한 그룹의 수는 1과 18 사이의 수로 대략적으로 2-4 사이로 보인다.

이상의 대략적인 자료탐색을 기반으로 이제 잠재그룹 포아송 모형을 사용하여 주어진 자료를 2-4 장에서 제안된 방법을 사용하여 분석하고자 한다. 분석의 목표는 사망자수를 중심으로 서로 동질적인 연령끼리 묶어 그룹화하고 각 그룹의 분포에 대한 모수를 추정하는 것이다. 아울러 적절한 그룹의 수를 통계적으로 추정하고자 한다.

5.2. 사후 추론

먼저 가능한 그룹의 수를 1,2,3,4로 하고 각 1-4의 그룹의 수를 갖는 잠재그룹 포아송 모형의 모수들을 추정한다. 즉, 네 개의 잠재그룹 포아송 모형을 추정하게 되는 것이다. 3장에서 제시된 각각의 모수들에 대한 사후 조건부 분포를 사용하여 깁스 샘플링을 적용하면 모수들의 사후표본들을 얻을 수 있다. 표본들의 time sequence plot을 관측한 결과 매우 빨리 수렴함을 알 수 있었으며 수렴시까지 걸리는 burn-in time 으로 3000개를 택하였다. 이 burn-in-time 이후에 총 10,000개의 모수에 대한 사후표본을 얻었고 이를 이용하여 모수의 추정치와 추정오차를 계산하였다. 또한 이 모수들의 추정치를 이용하여 각 객체, 즉, 각 연령의 그룹에 대한 소속확률을 계산하고 이 소속확률이 가장 큰 그룹을 객체가 실제로 속하는 소속그룹으로 추정하였다.

각 그룹 수에 대한 잠재그룹 포아송 모형의 모수 추정치와 소속그룹을 각각 표 5.2와 표 5.3에 나타내었다. 예를 들면, 2그룹-잠재 포아송 모형의 경우 60세 미만과 60세 이상의 두

연령그룹으로 분류하였고 이 두 그룹의 평균을 각각 3.24와 85.0으로 추정하였다. 3그룹-잠재 포아송 모형의 경우, 50세 미만, 50-64세, 65세 이상의 세 연령그룹으로 분류하였고 이들 그룹들이 각각 평균적으로 0.89, 22.08, 94.87의 사망자수를 갖는 것으로 추정하였다. 표 5.1에 주어진 자료를 살펴볼 때 이 분류는 상당히 합리적임을 알 수 있다.

표 5.2: 계층에 따른 모수들의 추정치

K	ϵ	λ
1	1	30.5
2	0.6504 0.3496	3.24 85.00
3	0.5222 0.1926 0.2852	0.89 22.08 94.87
4	0.4513 0.1560 0.1267 0.2659	0.47 9.72 33.64 95.61

5.3. 적절한 그룹수의 추정

5.2절의 결과는 그룹의 수가 고정된 경우에 한하였다. 1부터 4까지의 그룹수를 고려하였는데 4장에서 제시한 방법을 사용하여 어떤 그룹수가 주어진 자료에 가장 적절한지 추정하고자 한다.

4장에 따르면 $\pi(y|K)$ 혹은 $log\pi(y|K)$ 값을 가장 크게 하는 K값이 LC 모형에서 가장 적절한 그룹의 수이다. 표 5.4는 4장에서 제시한 방법으로 $log\pi(y|K)$ 값을 구한 것이고, 그림 5.3은 표 5.4의 값을 그래프로 나타낸 것이다. 이들 결과를 살펴보면, $K = 3$까지는 $log\pi(y|K)$가 급격히 증가하나 $K = 3$과 $K = 4$에서는 거의 비슷함을 알 수 있다. 즉, 연령별로 전립선암 사망자 수를 나타내는 이 자료에서는 3 또는 4그룹 포아송 모형이 가장 적절하다는 것을 의미한다. $K = 3$과 4에서 $logf(y|K)$값이 거의 동일하고 $K = 4$의 경우 소그룹이 존재하여 우리는 $K = 3$을 이상적인 그룹수로 정하기로 한다. 참고로 $K = 5$의 경우에는 빈그룹, 즉, 소속객체의 수가 0인 그룹이 존재하여 고려대상에서 제외하였다.

그림 5.2는 18개 사망자수 관측치에 대한 상대돗수 막대그림이다. 또한 겹쳐져 있는 곡선은 5.2절에서 추정된 3-그룹 잠재그룹 포아송 모형의 모수 추정치를 사용하여 추정된 확률밀도함수이다. 자료의 상대돗수 막대그림을 보면 3-그룹 모형이 적절함을 보여주고 있으며 추정된 밀도함수와도 매우 근접하여 본 논문에서 수행된 모수추정과 잠재그룹의 수 추정이 매우 적절함을 보여주고 있다. 3-그룹 모형에서 그룹 1은 약 52의 연령을 포함하여 크기는 가장 크나 평균은 0.89로 가장 작은, 젊은 층을 포함하는 그룹이며, 그룹 2는 크기가 가장 작고 평균은 22.08로 중간에 위치하며, 그룹 3은 평균이 94.87로 가장 큰, 즉, 가장 큰 사망자수를 갖는 노년층 그룹임을 알 수 있다. 각 연령그룹에 대하여 소속확률을 조사하니 표 5.3의 분류에 따라 소속그룹의 확률이 거의 모두 1에 가까웠다. 따라서 이 자료는 명확하게 그룹화할 수 있는 것으로 보인다.

표 5.3: LC 모형에서 각각의 객체들이 속하는 계층

obs.	연령별	사망자수	$K=1$	$K=2$	$K=3$	$K=4$
1	1-4세	0	1	1	1	1
2	5-9세	0	1	1	1	1
3	10-14세	0	1	1	1	1
4	15-19세	0	1	1	1	1
5	20-24세	0	1	1	1	1
6	25-29세	0	1	1	1	1
7	30-34세	1	1	1	1	1
8	35-39세	2	1	1	1	1
9	40-44세	0	1	1	1	1
10	45-49세	5	1	1	1	2
11	50-54세	10	1	1	2	2
12	55-59세	20	1	1	2	3
13	60-64세	36	1	2	2	3
14	65-69세	78	1	2	3	4
15	70-74세	100	1	2	3	4
16	75-79세	113	1	2	3	4
17	80-84세	104	1	2	3	4
18	85세 이상	79	1	2	3	4

표 5.4: 계층의 수에 따른 $log\pi(y|K)$ 값

K	$log\pi(y\|K)$
1	-534.1
2	-122.1
3	-82.4
4	-78.4

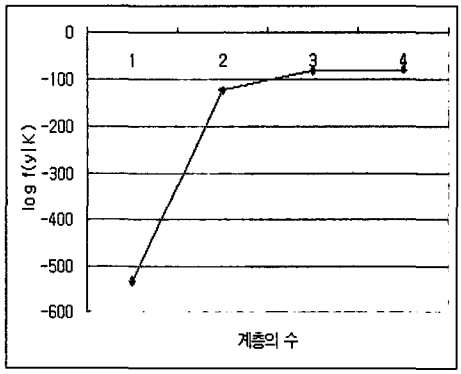

그림 5.3: 계층의 수에 따른 $log\pi(y|K)$ 값

6. 결 론

본 논문에서는 사망원인 통계 연보에 주어진 2000년도 연령별 전립선암 사망자수에 대하여 잠재그룹 포아송 모형을 가정하고 이에 대한 베이지안 그룹화 방법을 제안하였다. 분석 결과 18개의 연령그룹 대신 1-50세, 50-64세, 65세 이상의 세 그룹으로 연령을 구분하는 것이 적절함을 알 수 있었고 이러한 3-그룹 잠재그룹 포아송 모형이 자료에 잘 적합함을 보여주었다. 또한 각 연령 마다 각 그룹에 대한 소속확률을 제공하여 두 그룹 사이의 중간에 위치한 연령과 뚜렷이 어느 한 그룹에 속하는 연령들을 구분할 수 있게 하였다. 이러한 연령들의 그룹화 및 그룹의 특성에 대한 추정은 병에 대한 실제 역학조사 및 병의 관리에 있어서 실제적으로 큰 도움이 될 것이다.

본 논문에서는 2000년도의 자료만 분석하였으나 통계연보에는 1983-2000년의 자료가 제공되고 있다. 연령별 그룹화 외에 년도별 그룹화를 할 수 있다면 병의 추세가 시간에 따라 어떻게 달라지는지에 대한 정보를 제공하여 공중보건에 공헌을 할 수 있을 것으로 생각된다. 즉, 어느 시기에 병의 패턴이 이전의 것과 유의하게 달라지는지 알 수 있다면, 예를 들면 젊은 층의 암 사망자수의 증가 등, 그 시기에 이루어진 생활양식 및 식생활의 변화와 병과의 관계를 유추하는데 도움이 될 것이다. 이를 위해서는 두개의 변수를 고려한 포아송 모형의 설정이나 연도별로 연령별 자료를 효과적으로 요약하는 방법의 개발이 필요하며 추후로 이에 대한 보다 심층적인 연구가 요구된다고 본다.

잠재그룹 모형에 대한 베이지안 추정과 잠재그룹 수를 결정하는 방법으로 역점프 MCMC (reversible jump MCMC)기법의 적용도 가능하다. 그러나 역점프 MCMC 는 적용이 매우 복잡하며 서로 다른 K 값에 대응하는 모형간의 이동시에 표본생성함수로 사용되는 proposal density 의 선택이나 모형간의 이동순서에 따라 효율이 영향을 받는다는 단점이 있다. 반면에 본 논문에서 제안된 기법은 이미 추정을 위해 생성된 MCMC 표본을 사용하여 잠재그룹의 수를 결정하는 통계량을 계산하므로 경제적이며, 깁스표본기법을 사용하므로 주관적인 proposal density를 선택할 필요가 없어 비전문가도 쉽게 사용항 수 있다는 장점이 있다.

참고문헌

DeSarbo, W.S. and Choi, J. (1999). A latent structure double hurdle regression model for exploring heterogeneity in consumer search patterns, *Journal of Econometrics*, **89**, 423-455.

DeSarbo, W.S. and Cron, W.L. (1988). A maximum likelihood methodology for clusterwise linear regression, *Journal of Classification*, **5**, 249-282.

DeSoete and DeSarbo, W.S. (1991). A latent class probit model for analyzing pick any/n data, *Journal of Classification*, **8**, 45-64.

Forman, A.K. (1985). Constrained latent class models : theory and applications, *British Journal of Mathematical Statistics and Psychology*, **38**, 87-111.

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.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, **61**, 215-231.

Heinen, T. (1993). *Discrete Latent Variable Models*, Tilberg University Press.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values : application to educational testing, *Statistica Sinica*, **8**, 691-711.

Jedidi, K., Ramaswamy, V. and DeSarbo, W.S. (1993). A maximum likelihood method for latent class regression involving a censored dependent variable, *Psychometrika*, **58**, 375-394.

Kamakura, W.A. and Russel, G. (1989). A probabilistic choice model for market segmentation and elasticity structure, *Journal of Marketing Research*, **26**, 379-390.

Oh, M.S. (1999). Estimation of posterior density functions from a posterior sample, *Computational Statistics and Data Analysis*, **29**, 411-428.

[2004년 4월 접수, 2004년 8월 채택]

Bayesian Clustering of Prostate Cancer Patients by Using a Latent Class Poisson Model *

Man-Suk Oh [1)]

ABSTRACT

Latent Class model has been considered recently by many researchers and practitioners as a tool for identifying heterogeneous segments or groups in a population, and grouping objects into the segments. In this paper we consider data on prostate cancer patients from Korean National Cancer Institute and propose a method for grouping prostate cancer patients by using latent class Poisson model. A Bayesian approach equipped with a Markov chain Monte Carlo method is used to overcome the limit of classical likelihood approaches. Advantages of the proposed Bayesian method are easy estimation of parameters with their standard errors, segmentation of objects into groups, and provision of uncertainty measures for the segmentation. In addition, we provide a method to determine an appropriate number of segments for the given data so that the method automatically chooses the number of segments and partitions objects into heterogeneous segments.

Keywords: Latent class model, Mixture model, Markov chain Monte Carlo, Model selection.

* This work was supported by grant No. (R06-2002-012-01002-0) from the Basic Research Program of the Korea Science and Engineering Foundation

1) Professor, Department of Statistics, Ewha Women's University, 21 Daehyun-dong , Sodaemoon Gu, Seoul 120-750, Korea
E-mail: msoh@mm.ewha.ac.kr