응용통계연구 제 18권 2호, 2005년, pp. 471–485

RAINDROP PLOT을 이용한 차원축소

홍종선[1] 김범준[2] 박지용[2]

요 약

범주형 자료분석에서 차원축소(collapsibility)는 오즈비로 설명되었다. 실제의 2×2×K 분할표 자료를 이 이론에 적용시켰을 때 오즈비의 값으로 차원축소가 가능한지의 여부를 판단하기는 어렵다. 오즈비를 시각적으로 표현하는 방법 중에서 Doi, Nakamura와 Yamamoto(2001)가 제안한 Contour plot을 통해서 분할표 자료를 설명하는 것은 가능하지만 차원축소의 가능성을 결정하기에는 한계가 있다. 본 연구에서는 오즈비의 신뢰구간을 시각적으로 표현할 수 있는 방법으로 Barrowman과 Myers(2003)가 제안한 Raindrop plot을 이용하여 $2 \times 2 \times K$ 분할표 자료를 설명할 수 있으며 동시에 차원축소의 가능성을 판단할 수 있는 방법을 제안하고자 한다.

주요용어: 오즈비, 교차적비, 차원축소, 로그선형모형

1. 서론

이차원 평면에 범주형 자료를 표현하는 다양한 시각적인 방법은 자료구조를 이해하기 위해 유용한 정보를 제공하는 탐색적 자료 분석(EDA)을 수행한다. 잘 알려진 EDA에 의한 시각적인 방법은 히스토그램(histogram), 바차트(bar chart), 파이차트(pie chart), 스타차트(star chart) 등이 있다. 이것들은 하나의 범주형 변수에 대해 다양한 범주들의 확률 또는 빈도를 보여준다. Fienberg(1975, 1980)는 이차원 범주형 자료 중 2×2 분할표를 표현하는 'four-fold circuit display'를 제안하였고, 이것은 각 칸의 빈도에 비례하는 반지름을 갖는 4개의 원으로 구성되어 있다. 이차원 분할표를 표현하는 대표적인 방법으로는 블록차트(block chart)가 있으며, Hartigan과 Kleiner(1981, 1984)는 타일(tile)이라 불리는 정사면체의 크기를 관찰값의 확률에 비례하도록 표현한 'mosaic plot'을 제안하였다. 독립 모형하에서 Cohen(1980)과 Friendly(1992)는 각 칸의 편차를 표현한 'association plot'을 제안하였다. Friendly(1992, 1994)는 각 타일에 피어슨 카이제곱의 각 칸에 해당하는 편차의 크기를 고려하여 색상과 빗금으로 표현한 보다 향상된 'mosaic plot'을 제안하였다. Tukey(1977)는 이차원 분할표에 대한 적합도를 표현한 'two-way plot'을 제안하였다.

$I \times J \times K$ 삼차원 자료에 대해 세 번째 변수의 각 범주와 분리된 $I \times J$ 분할표는 'mosaic plot'과 'four-fold circular display'로 분석된다. 사차원 또는 그 이상의 고차원 분할표는 향상된 'mosaic plot'을 통해 표현할 수 있다(Hartigan과 Kleiner 1984). 오민권, 홍종선과 이종철(1999)은 다차원 분할표 자료를 분석하고 이를 시각적인 방법으로 설명한 'ring chart'

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-mail: cshong@skku.ac.kr
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원생

방법을 제안하였다. 이는 모든 칸의 확률을 링모양의 그림으로 보여주어 범주형 자료의 전체 구조를 설명할 수 있게 해주며 그 외에 'standardized ring chart', 'double ring chart', 'residual ring chart'를 제안하였다. 이 방법들은 주어진 로그선형모형의 적합도를 평가하는 것뿐만 아니라 일차, 이차교호작용을 포함하는 변수들 사이의 관계를 결정하는데 유용한 정보를 얻을 수 있다(홍종선과 이종철 2000).

Fienberg(1970), Fienberg와 Gilbert(1970)는 2×2 분할표에 대해 사면체 내의 궤적(loci)을 환산하여 변수간의 연관성 측도를 기하학적으로 표현하는 방법을 제안하였다. Darroch, Lauritzen와 Speed(1980)는 다차원 분할표에 대한 독립모형과 조건부 독립모형을 표현할 수 있는 그래픽 모형(graphical model)을 개발하였다. 이러한 그래픽 모형들은 변수들 사이에 연관성의 측도를 나타내는 연관그래프(association graph)로 표현하였다.

자료 전체 구조를 설명해주고, 변수들 사이의 관계를 분석하는데 도움을 주는 여러 시각적인 방법 이외에 오즈비(odds ratio) 또는 교차적비(cross product ratio)를 이용하여 다양하게 시각적으로 자료의 구조를 설명하는 방법들이 있다. 그 중에서 Doi, Nakamura와 Yamamoto(2001)가 제안한 Contour plot은 $2 \times 2 \times K$ 분할표 자료의 설명은 가능하지만 차원축소(collapsibility) 가능성을 결정하기에는 한계가 있다. 실제의 $2 \times 2 \times K$ 분할표 자료를 오즈비의 값으로 차원축소가 가능한지를 판단하기는 어렵기 때문에 본 연구에서는 오즈비의 신뢰구간을 시각적으로 표현할 수 있는 Barrowman과 Myers(2000, 2003)가 제안한 Raindrop plot을 이용하여 $2 \times 2 \times K$ 분할표 자료를 설명하고자 하며 동시에 차원축소 가능성을 판단할 수 있는 방법을 제안하고자 한다. 2절에서 Contour plot과 Raindrop plot을 간략히 설명하고, 3절에서는 차원축소가능모형과 불가능모형을 따르는 자료를 생성하고 생성된 자료를 Contour plot과 Raindrop plot으로 구현하여 차원축소에 대하여 논의하고자 한다. 4절에서 실제의 자료를 실증적으로 설명하였으며, 본 연구의 결론은 5절에서 유도하였다.

2. Contour plot과 Raindrop plot

우선 삼차원 분할표에서 기본이 되는 $2 \times 2 \times K$ 분할표에 대한 기본적인 개념과 수식을 정의하기로 한다. 아래의 표 2.1과 같이 분할표에서 칸의 빈도 x_{ijk}와 이에 대응하는 칸 확률 p_{ijk}를 가지는 $2 \times 2 \times K$ 분할표와 세 번째 변수에 대하여 차원축소된 2×2 분할표를 고려하자.

표 2.1: $2 \times 2 \times K$ 분할표와 차원축소된 2×2 분할표

p_{111}	p_{121}	p_{112}	p_{122}	...	...	p_{11K}	p_{12K}
p_{211}	p_{221}	p_{212}	p_{222}			p_{21K}	p_{22K}

p_{11+}	p_{12+}
p_{21+}	p_{22+}

임의의 $k = 1, 2, \ldots, K$에 대한 p_k, q_k를 다음과 같이 정의한다.

$$p_k = \frac{p_{11k}}{p_{+1k}}, \quad q_k = \frac{p_{12k}}{p_{+2k}}. \tag{2.1}$$

차원축소된 2×2 분할표에 대한 p_c, q_c를 다음과 같이 정의한다.

$$p_c = \frac{p_{11+}}{p_{+1+}}, \quad q_c = \frac{p_{12+}}{p_{+2+}}.$$

표 2.1에서 k번째 2×2 분할표의 교차적비와 차원축소된 분할표의 교차적비는 역시 다음과 같이 표현된다.

$$\theta_k = \frac{p_k/(1-p_k)}{q_k/(1-q_k)}, \quad \theta_c = \frac{p_c/(1-p_c)}{q_c/(1-q_c)}. \tag{2.2}$$

함수 $f(p, q)$와 그 함수값 θ에 대응하는 Contour $C(\theta)$를 정의하면 다음과 같다.

$$f(p,q) = \frac{p/(1-p)}{q/(1-q)}, \quad (p,q) \in (0,1)^2,$$

$$C(\theta) = \left\{ (p,q) \in (0,1)^2 \ : \ f(p,q) = \theta \right\}, \ \theta > 0.$$

그림 2.1은 $\theta = 0.1, 0.5, 1, 2$와 10에 대한 Contour $C(\theta)$의 형태를 나타낸다.

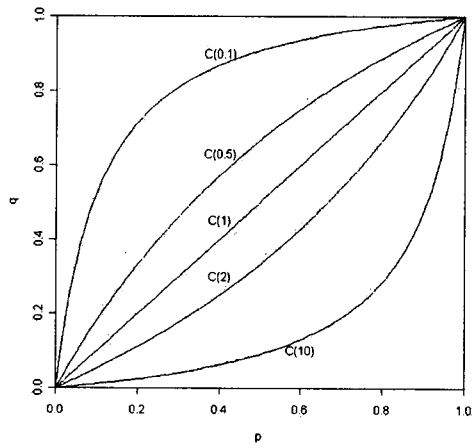

그림 2.1: $C(\theta)$의 형태

2×2 분할표의 주변합을 고정시켰을 때 로그오즈비 $\theta_k^l = log(\theta_k)$의 조건부 가능도함수는 다음과 같이 얻을 수 있다(Agresti 1990).

$$L(\theta_k^l) = \binom{x_{11k} + x_{21k}}{x_{11k}} \binom{x_{21k} + x_{22k}}{x_{21k}} e^{\theta_k^l x_{11k}} / S(\theta_k^l),$$

여기서 $S(\theta_k^l)$는 분할표의 주변합을 고정시켰을 때 가능한 x_{11k} 값에 대한 가능도함수의 분자의 총합이다.

로그오즈비 θ_k^l의 최대가능도추정량(MLE)을 θ_k^{MLE}라 표시하고, $l(\theta_k^l)$를 로그조건부 가능도함수로 정의하면, 가능도비의 근사분포에 기초하여 θ_k^l에 대한 근사적인 $100\times(1-\gamma)\%$ 신뢰구간은 다음과 같이 정의된다(Cox와 Hinkley(1974) 참조).

$$\{\theta_k^l : 2[l(\theta_k^{MLE}) - l(\theta_k^l) \le \chi^2_{1(1-\gamma)}]\},$$

여기서 $\chi^2_{1(p)}$는 자유도 1을 갖는 카이제곱분포의 $100\times p$ 번째 백분위수이다. 편의상 $l(\theta_k^{MLE})$ = 0이라 설정하면, $\chi^2_{1(1-\gamma)}/2$이기 때문에 $100\times(1-\gamma)\%$ 신뢰구간은 다음과 같다.

$$\{\theta_k^l : l(\theta_k^l) \ge -\chi^2_{1(1-\gamma)}/2\}.$$

$\gamma = 0.05$일때 $\chi^2_{1(0.95)}/2 = 1.92$이므로, Raindrop plot은 로그가능도 함수가 -1.92보다 큰 부분을 반사시켜 작성한다. Raindrop plot의 구현은 Barrowman과 Myers(2003)가 S+/R로 작성된 프로그램을 사용하며 그림 2.2는 2×2 분할표자료를 Raindrop plot으로 구현한 예이다. 두 개의 물방울 모양을 한 Raindrop plot 중 안쪽의 Raindrop은 95% 신뢰구간을 그리고 바깥쪽은 99% 신뢰구간을 표현하였다.

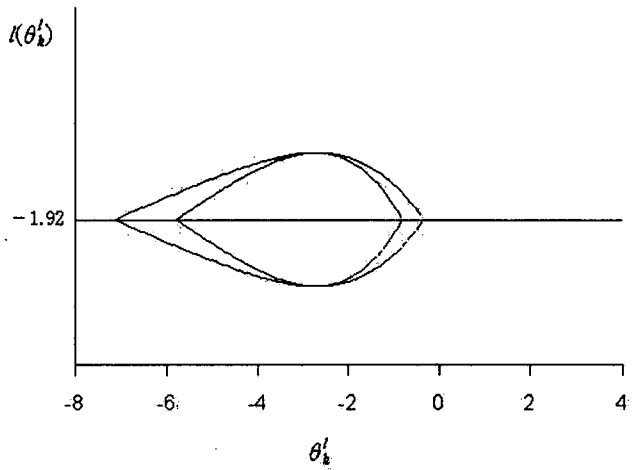

그림 2.2: Raindrop plot

Raindrop plot에서는 로그오즈비의 신뢰구간을 표현하였기 때문에 신뢰구간에 0값이 포함여부를 파악할 수 있어 유용하다. 즉 Raindrop plot에서 로그오즈비의 신뢰구간에 0값이 포함된다면, 두 범주형 변수가 동질 또는 독립이라고 설명할 수 있기때문에 Contour plot에서의 Contour가 직선 또는 곡선 형태인지 판단하는 것보다 통계적으로 결정내릴 수 있는 장점이 있다.

3. 차원축소와 Raindrop plot

삼차원 범주형 자료에 대한 로그선형모형들을 Christensen(1990)이 사용한 표기 방법을 따라 분류하면 다음과 같다: 포화모형(saturated model)은 [123], 부분연관모형(partial association model)은 [12][13][23], 조건부독립모형(conditionally independent model)은 [12][13], [12][23], [13][23], 한 변수의 독립모형(model with one factor independent of the other two)은 [12][3], [13][2], [1][23], 그리고 완전독립모형(completely independent model)은 [1][2][3]으로 표기한다.

Bishop, Fienberg와 Holland(1975, pp. 39, 47)는 삼차원 분할표에서 한 변수가 다른 두 변수와 조건부독립이기만 하면 한 변수의 주변합으로 구성된 다른 두 변수간의 주변표로의 차원축소가 가능하다고 정의하였는데, 세 번째 변수를 교락(confounder)변수로 간주하여 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형은 [12][3], [12][13], [12][23]이며 그 이외 다른 모형은 차원축소 불가능한 모형이다(자세한 이론은 Agresti(1984, pp. 146), 그리고 Christensen(1990, pp. 114)을 참조할 것).

Ducharme과 Lepage(1986)는 강한 차원축소(strong collapsibility)의 필요충분조건을 다음과 같이 정의하였다.

$$p_{ijk} = p_{ij+}p_{++k}, \quad p_{ijk} = p_{ij+}p_{i+k}/p_{i++}, \quad p_{ijk} = p_{ij+}p_{+jk}/p_{+j+}.$$

위 식의 세가지 조건을 만족하는 각각의 모형은 [12][3], [12][13], [12][23] 모형으로 재표현되는데 이들 모형들은 앞에서 언급한 바와 같이 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형이다. 따라서 차원축소 가능한 모형은 Ducharme과 Lepage(1986)이 정의한 강한 차원축소 가능한 모형과 동일하게 간주할 수 있다.

우선, 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형의 경우 (2.1) 식에서 정의한 p_k, q_k를 구하면 다음과 같다.

$$\left.\begin{aligned} p_k &= \frac{p_{11k}}{p_{+1k}} = \frac{p_{11+}p_{++k}}{p_{+1+}p_{++k}} = \frac{p_{11+}}{p_{+1+}} = p_c \\ q_k &= \frac{p_{12k}}{p_{+2k}} = \frac{p_{12+}p_{++k}}{p_{+2+}p_{++k}} = \frac{p_{12+}}{p_{+2+}} = q_c \end{aligned}\right\} \quad \text{[12][3] 모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{11+}p_{+1k}/p_{+1+}}{p_{+1+}p_{+1k}/p_{+1+}} = \frac{p_{11+}}{p_{+1+}} = p_c \\ q_k &= \frac{p_{12+}p_{+2k}/p_{+2+}}{p_{+2+}p_{+2k}/p_{+2+}} = \frac{p_{12+}}{p_{+2+}} = q_c \end{aligned}\right\} \quad \text{[12][23] 모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{11+}p_{+1k}/p_{1++}}{(p_{11+}p_{1+k}/p_{1++})+(p_{21+}p_{2+k}/p_{2++})} \\ q_k &= \frac{p_{12+}p_{1+k}/p_{1++}}{(p_{12+}p_{1+k}/p_{1++})+(p_{22+}p_{2+k}/p_{2++})} \end{aligned}\right\} \quad \text{[12][13] 모형}$$

[12][3], [12][23] 모형에서 p_k, q_k는 k에 의존하지 않으므로 각각 p_c, q_c와 동일하다. 따라서 모든 k에 대하여 $\theta_k = \theta_c$이다. 또한 [12][13] 모형에서 p_k, q_k는 k에 독립적이지 않지만 (2.2) 식에서 정의한 $\theta_k = p_{11+}p_{22+}/p_{12+}p_{21+}$가 되어 k와 독립이며 따라서 θ_c의 값과 동일함을 유도할 수 있다. 그러므로 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형에서는 모든 k에 대하여 $\theta_k = \theta_c$임을 얻을 수 있다.

다음으로 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 불가능한 모형([1][2][3], [1][23], [13][2], [13][23], [12][13][23]) 중에서 부분연관모형인 [12][13][23] 모형에서는 p_{ijk}에 대한 해를 직접적인 방법으로 구할 수 없기 때문에 p_k, q_k를 정의할 수 없다. 그러나 [1][2][3], [1][23], [13][2], [13][23] 모형에 대하여 p_k, q_k를 구하면 다음과 같다.

$$\left.\begin{aligned} p_k &= \frac{p_{1++}p_{+1+}p_{++k}}{p_{+++}p_{+1+}p_{++k}} = p_{1++} \\ q_k &= \frac{p_{1++}p_{+2+}p_{++k}}{p_{+++}p_{+2+}p_{++k}} = p_{1++} \end{aligned}\right\} \quad [1][2][3] \text{ 모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{1++}p_{+1k}}{p_{+++}p_{+1k}} = p_{1++} \\ q_k &= \frac{p_{1++}p_{+2k}}{p_{+++}p_{+2k}} = p_{1++} \end{aligned}\right\} \quad [1][23] \text{ 모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{1+k}p_{+1+}}{p_{++k}p_{+1+}} = \frac{p_{1+k}}{p_{++k}} \\ q_k &= \frac{p_{1+k}p_{+2+}}{p_{++k}p_{+2+}} = \frac{p_{1+k}}{p_{++k}} \end{aligned}\right\} \quad [13][2] \text{ 모형}$$

$$\left.\begin{aligned} p_k &= \frac{p_{1+k}p_{+1k}/p_{++k}}{p_{++k}p_{+1k}/p_{++k}} = \frac{p_{1+k}}{p_{++k}} \\ q_k &= \frac{p_{1+k}p_{+1k}/p_{++k}}{p_{++k}p_{+1k}/p_{++k}} = \frac{p_{1+k}}{p_{++k}} \end{aligned}\right\} \quad [13][23] \text{ 모형}$$

차원축소 불가능한 모형 중 [1][2][3], [1][23], [13][2], [13][23] 모형에 대한 p_k, q_k는 모두 동일하기 때문에 모든 오즈비의 값은 1이다. 즉 모든 k에 대하여 $\theta_k = \theta_c = 1$이다.

$2 \times 2 \times K$ 분할표 자료에 대하여 위에서 연구한 오즈비에 관하여 정리하면 다음과 같다. 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 [12][3], [12][13], [12][23] 모형에서는 다음과 같은 관계식을 얻는다.

$$\theta_1 = \cdots = \theta_K = \theta_c \neq 1. \tag{3.1}$$

첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 불가능한 모형 중 [1][2][3], [1][23], [13][2], [13][23] 모형에서는 다음과 같은 관계를 유도하였다.

$$\theta_1 = \cdots = \theta_K = \theta_c = 1. \tag{3.2}$$

그리고 [12][13][23] 모형에서는 이차교호작용항이 존재하지 않기 때문에 오직 다음과 같은 관계식만을 유도할 수 있다.

$$\theta_1 = \cdots = \theta_K. \tag{3.3}$$

3절의 처음에 언급한 여러 종류의 로그선형모형에 적합한 3차원 분할표 자료 중에서 $2 \times 2 \times 3$ 자료를 생성하여 부록 B에 수록하였다(단순한 완전독립모형은 제외). 부분연관모형([12][13][23]) 이외의 모형은 모두 직접해(direct solution)을 구할 수 있는 모형들이며, 각 모형에 적합한 칸 확률 p_{ijk}는 주어진 충분합 형태(sufficient configuration)의 함수로 구하고 대응하는 칸 빈도 x_{ijk}는 표본크기 $N = 1{,}000$과 모비율 $\{p_{ijk}\}$로 이루어진 다항분포를 따르는 난수를 생성하는 모의실험을 통하여 구한다. 예를 들어, 조건부 독립모형 중

[13][23] 모형의 칸 확률 p_{ijk}는 주어진 충분합 형태 $\{p_{i+k}\}$, $\{p_{+jk}\}$의 주변확률표를 이용하여 $p_{ijk} = p_{i+k}p_{+jk}/p_{++k}$의 관계식을 이용하여 구한다. 직접해가 존재하지 않은 부분연관모형인 경우에는 정동빈, 홍종선과 윤상호(2003)의 연구에서 사용한 방법을 사용하여, 적절한 구간의 균일분포를 따르는 난수를 생성하고 로그선형모형의 여러 모수의 추정값을 얻은 후 이에 대응하는 칸 빈도 x_{ijk}를 생성하였다.

부록 B에 수록된 $2 \times 2 \times 3$ 분할표 자료에 대하여 Contour plot과 Raindrop plot을 작성하여 부록 A의 그림 A-1부터 그림 A-7에 수록하였다. 우선 각 Contour plot에서 4개의 Contour 중 3개의 Contour는 원 분할표자료에서 3개의 2×2 분할표에 대한 것이다. 각 p_k, q_k에 대응하는 좌표에 원모양의 점으로 표현하였고, 나머지 하나의 Contour는 차원축소된 2×2 자료에 대응하는 것으로 p_c, q_c의 좌표는 속이 채워진 원모양으로 표시하였다. Raindrop plot에서도 4개의 Raindrop이 존재하는데 마지막 하단부의 Raindrop이 차원축소된 분할표에 대한 것이다.

부록 B의 자료 이외에 여러 종류의 로그선형모형에 적합한 자료를 여러 번 생성하고 대응하는 Contour plot과 Raindrop plot을 작성하여 살펴본 결과 다음과 같은 결론을 유도할 수 있었다. 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형에서는 (3.1) 식에 정리된 이론과 같이 모든 오즈비 θ_k, θ_c 값은 유사하며 모두 1에서 멀리 떨어진 값을 갖는다. 그림 A.1부터 그림 A.3까지의 Contour plot에서 모든 Contour들은 좌표 (0,0)과 (1,1)을 연결하는 대각선에서 멀리 떨어진 오목과 볼록 형태로 표현되고 있으며, Raindrop plot에서는 로그오즈비의 95%와 99% 신뢰구간에 0값을 포함하지 않는다는 것을 발견하였다. 따라서 이 모형들의 모든 오즈비 θ_k, θ_c 값이 1이 아님을 Raindrop plot을 통해서 통계적으로 확신할 수 있다.

첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 불가능한 모형 중 [1][2][3], [1][23], [13][2], [13][23] 모형에 대하여는 (3.2) 식에 정리된 이론을 그림 A.4, A.5, 그림고 A.6의 Contour plot으로 파악하고 Raindrop plot으로 확인할 수 있다(단순한 완전독립모형인 [1][2][3]모형에 대하여는 plot을 생략함). 즉 Contour plot에서는 모든 Contour들이 대각선 근처에 집중되고 있으며 Raindrop plot에서는 로그오즈비의 신뢰구간에 0값이 포함되어 있다.

또한 차원축소 불가능한 모형 중 이차교호작용항이 존재하지 않는 [12][13][23] 모형에서는 오즈비에 대하여 (3.3) 식 이외에 어떠한 이론을 유도할 수 없는데, 그림 A.7의 Contour plot과 Raindrop plot을 살펴보면 그림 A.4부터 그림 A.6까지와 모두 상이한 형태임을 발견할 수 있었다.

4. 결론

3절에서 토론한 여러 로그선형모형에 대하여 적합한 자료를 바탕으로 작성한 Contour plot과 Raindrop plot을 살펴보면, θ_k, θ_c에 대한 명확한 이론이 설정된 [12][3], [12][13], [12][23] 모형과 [1][2][3], [1][23], [13][2], [13][23] 모형에 대하여는 그림 A.1부터 그림 A.6까지의 결과와 일치하는 것을 살펴볼 수 있다. 즉 [12][3], [12][13], [12][23] 모형에서는 모든 오

즈비의 값이 동일하고 로그오즈비의 신뢰구간에 0값이 포함되지 않았으며, [1][2][3], [1][23], [13][2], [13][23] 모형에서는 모든 오즈비의 값이 1에 가까운 것을 Contour plot을 이용하여 파악할 수 있으며, Raindrop plot을 이용해서는 통계적으로 확신할 수 있었다.

본 연구에서는 Ducharme과 Lepage(1986)에서는 강한 차원축소(strong collapsibility)의 정의와 Christensen(1990, pp. 113), Agresti(1984, pp. 146)가 내린 일반적인 차원축소에 대한 정의를 따라서 우리는 다음과 같은 결론을 유도할 수 있으며, 이와 같은 판단 기준은 부록 A의 그림 A.1부터 그림 A.6에서와 같이 Contour plot보다는 Raindrop plot을 사용하여 보다 통계적으로 설정할 수 있다.

- 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형은 강한 차원축소 모형으로 정의하는데 이 모형에 대한 Contour plot에서는 좌표 (0,0)과 (1,1)을 연결하는 대각선과는 멀리 떨어진 곳에 Contour가 위치하는데 이 Contour 값이 1이 아니다는 사실은 로그우도비의 신뢰구간에 0값을 포함하지 않는다는 것을 시각적으로 보여준 Raindrop plot을 통해서 통계적으로 설명이 가능하다. 따라서 Raindrop plot을 이용하여 세 번째 변수에 차원축소 가능한 모형의 모든 오즈비 θ_k와 θ_c의 값은 유사하며 그 값들이 1이 아니다라는 (3.1) 식의 이론과 일치함을 식별할 수 있다.

- 세 번째 변수가 차원축소 불가능한 모형 중 [1][2][3], [1][23], [13][2], [13][23] 모형에 대한 모든 Contour plot에서는 대각선과 가까이에서 Contour가 집중되는데 이 Contour 값이 1에 근접하다는 사실은 로그우도비의 신뢰구간에 0값을 포함한다는 것을 시각적으로 보여준 Raindrop plot을 통해서 통계적으로 설명이 가능하다. 따라서 Raindrop plot을 이용하여 이 모형의 모든 오즈비 θ_k와 θ_c의 값은 모두 1에 근접한 유사한 값을 갖고 있음을 식별할 수 있으며 이것을 요약한 관계이론은 (3.2) 식과 같음을 발견하였다.

참고문헌

Agresti, A. (1984). *Analysis of Ordinary Categorical Data*, John Wiley and Sons.

Agresti, A. (1990). *Categorical Data Analysis*, John Wiley and Sons.

Barrowman, N. J. and Myers, R. A. (2000). Still more Spawner-recruitment curves: The Hockey Stick and Its generalizations, *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 665-676.

Barrowman, N. J. and Myers, R. A. (2003). Raindrop plots: A new way to display collections of likelyhoods and distributions, *The American Statistician*, **57**, 268-274.

Bishop, Yvonne M. M., Fienberg, Steve E., and Holland, Paul W. (1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.

Christensen, Ronaldo. (1990). *Log-Linear Models*, New York: Springer-Verlag.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contigency table, *Communications in Statistics - Theory and Methods*, A9, 1025-1041.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman Hall.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables, *Annals of Statistics*, **57**, 552-539.

Doi, M., Nakamura, T., and Yamamoto, E. (2001). Conservative tendency of the crude odds ratio, *Journal of Japan Statistical Society*, **1**, 1-19.

Ducharme, G. R. and Lepage, Y. (1986). Testing collapsibility in contingency tables, *Journal of the Royal Statistical Society*, B, **48**, 197-205.

Efron, B. (1996). Empirical Bayes methods for combining likelihoods, *Journal of the American Statistical Association*, **91**, 538-565.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of 2×2 contingency tables, *Journal of the American Statistical Association*, **65**, 694-701.

Fienberg, S. E. (1975). Perspective Canada as a social report, *Social Indicators Research*, **2**, 154-174.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Data*, 2nd ed, The MIT press.

Friendly, M. (1992). Mosaic displays for log-linear models, *Proceedings of the Statistical Graphics Section, the American Statistical Association*, 61-68.

Friendly, M. (1994). Mosaic displays for multi-way contigency tables, *Journal of the American Statistical Association*, **89**, 190-200.

Hartigan, J. A. and Kleiner, B. (1981). Mosaic for contingency tables, *Computer Science and Statistics : Proceedings of the* 13^{th} *Symposium on the Interface*, ED. W. F. Eddy, New York : Spring-Verlag, 268-273.

Hartigan, J. A. and Kleiner, B. (1984). A Mosaic of the television ratings, *The American Statstician*, **38**, 32-35.

Hong, C. S. and Lee, J. C. (2000). Ring chart II for multidimensional categorical data analysing using conditional ring charts, *Korean Journal of Applied Statistics*, **13**, 163-178.

Hyndman, R. J. (1996). Computing and graphing highest density regions, *The American Statistician*, **50**, 120-126.

Jeong, D. B., Hong, C. S., and Yoon, S. H. (2003). Empirical comparisons of disparity measures for partial association models in three dimensional contingency tables, *The Korean Communications in Statistics*, **10**, 135-144.

Oh, M. G., Hong, C. S., and Lee, J. C. (1999). Ring chart for categorical data, *Korean Journal of Applied Statistics*, **12**, 225-240.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley Publishing Company.

Yamamoto, E. and Doi, M. (2001). Noncollapsibility of common odds ratios without/with confounding, *Bulletin of The* 53^{rd} *Session of the International Statistical Institute*, Book 3, 39-40.

[2004년 8월 접수, 2005년 3월 채택]

부록 A:

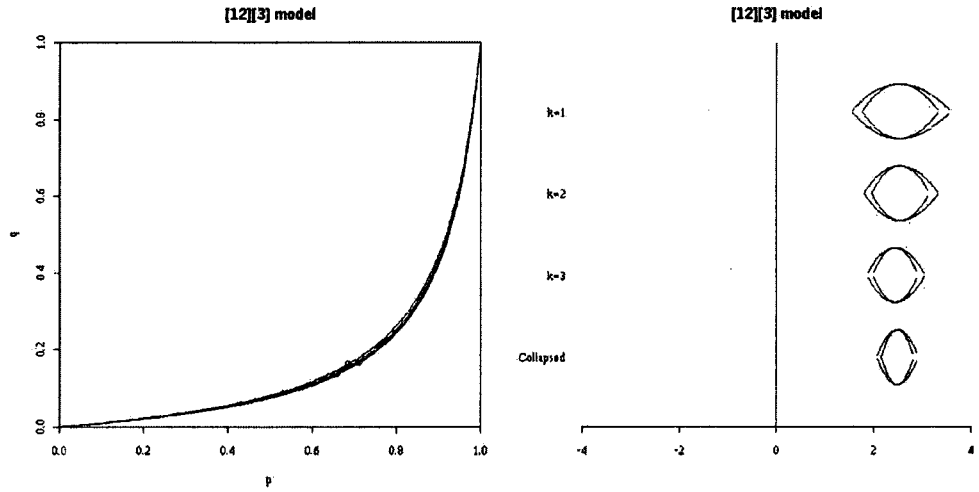

그림 A.1: 차원축소가능모형 [12][3]의 Contour plot과 Raindrop plot

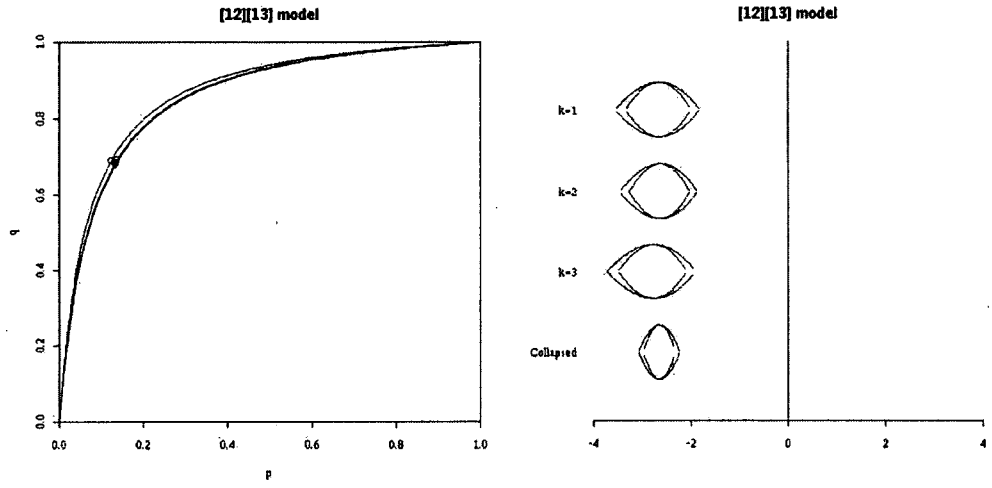

그림 A.2: 차원축소가능모형 [12][13]의 Contour plot과 Raindrop plot

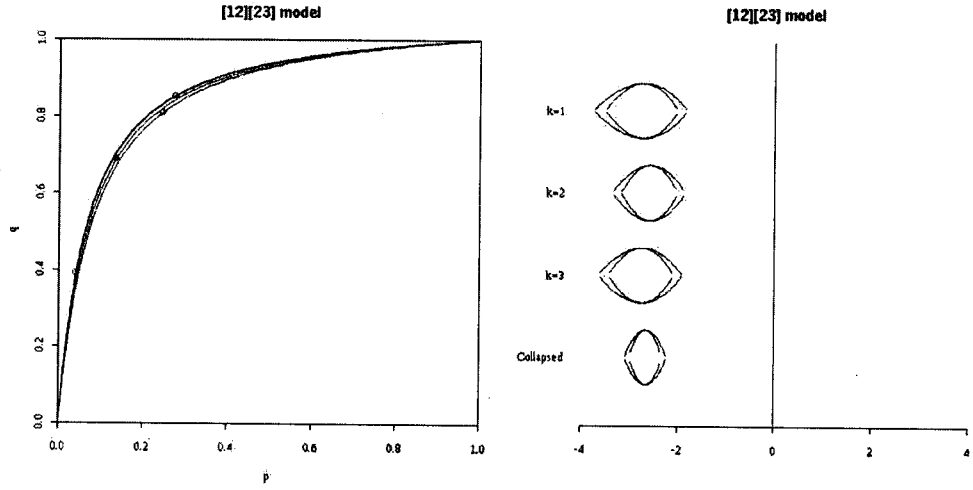

그림 A.3: 차원축소가능모형 [12][23]의 Contour plot과 Raindrop plot

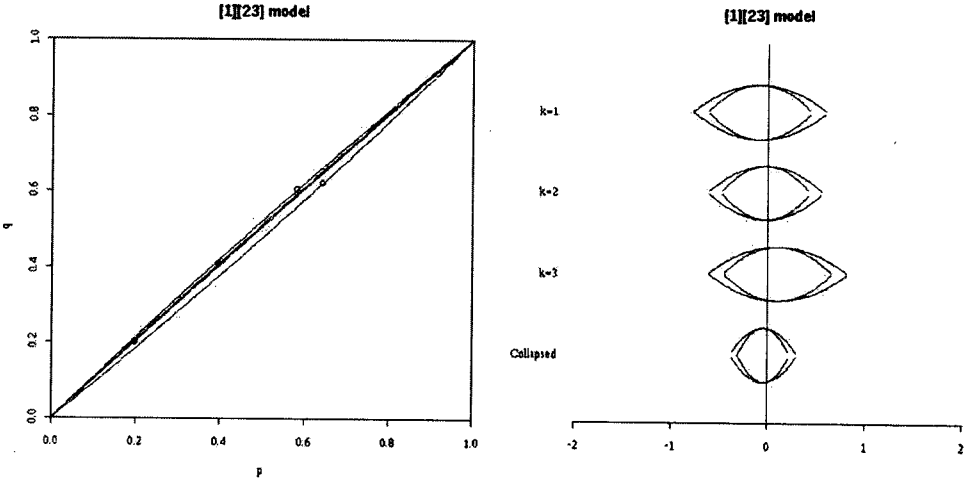

그림 A.4: 차원축소불가능모형 [1][23]의 Contour plot과 Raindrop plot

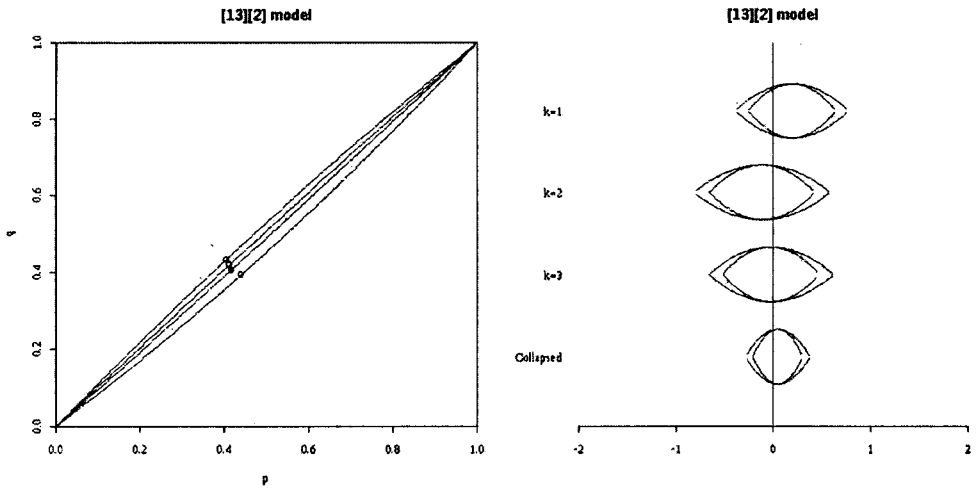

그림 A.5: 차원축소불가능모형 [13][2]의 Contour plot과 Raindrop plot

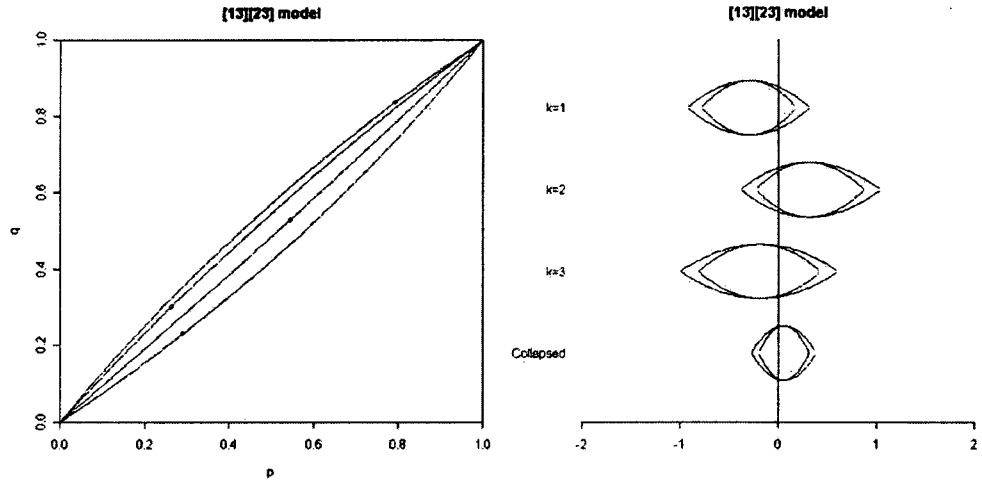

그림 A.6: 차원축소불가능모형[13][23]의 Contour plot과 Raindrop plot

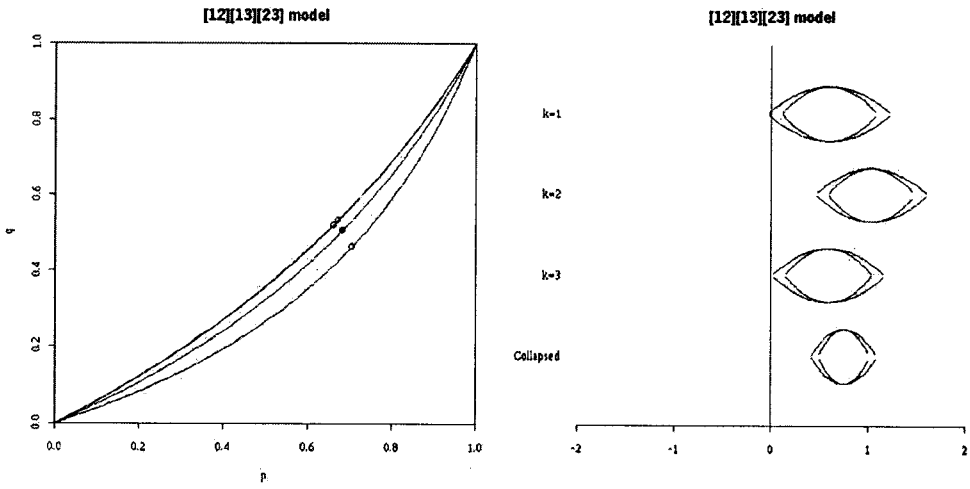

그림 A.7: 차원축소불가능모형[12][13][23]의 Contour plot과 Raindrop plot

부록 B:

[12][3]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	31	16	62	25	94	43
$i=2$	19	121	37	187	59	306

[12][13]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	59	372	17	111	14	100
$i=2$	36	16	108	51	80	36

[12][23]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	33	87	47	145	16	384
$i=2$	69	12	93	22	36	56

[1][23]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	53	38	38	155	56	31
$i=2$	96	63	63	251	97	59

[13][2]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	48	61	55	81	101	144
$i=2$	126	194	42	55	39	54

[13][23]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	236	62	54	132	25	70
$i=2$	163	32	26	87	34	79

[12][13][23]	$k=1$		$k=2$		$k=3$	
모형	$j=1$	$j=2$	$j=1$	$j=2$	$j=1$	$j=2$
$i=1$	106	54	164	68	136	66
$i=2$	64	59	64	74	78	68

Collapsibility Using Raindrop Plot

C. S. Hong[1] B. J. Kim[2] J. Y. Park[2]

ABSTRACT

For categorical data analysis, the collapsibility were explained with the odds ratio (cross-product ratio). When these theories with these odds ratios are applied to real $2 \times 2 \times K$ contingency tables, it is impossible to decide whether data are collapsible. Among graphical methods to represent odds ratios, Contour plot which is developed by Doi, Nakamura and Yamamoto (2001) could explain the structure of these data, but cannot decide on the collapsibility. In this paper, by using the Raindrop plot proposed by Barrowman and Myers (2003), we suggest an alternative method which can not only explain the structure of data, but also decide on the collapsibility.

Keywords: Odds ratio, Cross-product ratio, Collapsibility, Log-linear model.

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr

2) Graduate Student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.