응용통계연구 제 **20**권 3호, *2007*년, *pp. 585–595*

다차원 범주형 자료의 변환과 그의 응용*

안주선[1)]

요 약

Ahn 등 (2003)의 P-행렬을 사용한 두 c^d-분할표의 변환자료들의 유클리드 거리제곱은 두 분할표의 셀 (cell) 상대도수벡터들 사이의 유클리드 거리제곱에 비례함을 보이고, PP-자료의 플롯을 현대시분석과 설문자료의 탐색에 사용하는 방법을 제안한다.

주요용어: P-행렬, PP-플롯, 현대시, 설문자료.

1. 서론

Ahn 등 (2003)은 Radon 변환의 원리를 사용하여 P-행렬 (Projection matrix)을 만들고 이 행렬을 사용하여 c^d-분할표 자료를 변환했다. 변환 값 (PP-값)은 $(c^d-1)/(c-1)$ 방향에서 얻어진다. 또한 Ahn 등 (2003)은 각 방향에 대한 PP-값들을 2차원 상에 플롯 (plot)하고 이 플롯을 분할표들의 순서화와 집락 등의 다차원 범주형 자료의 특성을 조사하는 사영탐색방법 (Projection Pursuit Method, 간단히 PPM)을 논했다. $(c^d-1)/(c-1)$개의 PP-값들을 원소로 하는 집합을 PP-자료라 한다. 본 논문에서는 c가 소수인 경우 두 c^d-분할표의 PP-자료들 사이의 거리제곱은 상대도수로 나타낸 두 c^d-분할표 자료 벡터의 거리제곱의 상수배 (c와 d에만 의존)임을 보인다. 이는 연속형 다변량자료의 변환함수인 Andrews' curve

$$f_x(t) = x_1\sqrt{2} + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) + \cdots, \quad -\pi \le t \le \pi$$

의 거리유지 성질에 대응한 것으로 볼 수 있다 (Andrews, 1972).

P-행렬을 사용한 사영탐색 방법 (PPM)은 c^d-분할표로부터 서로 직교 (orthogonal)한 $(c^d-1)/(c-1)$개의 $d-1$차원 분할표상의 셀 (cell) 상대도수들의 합을 구하고 그 들의 플롯을 조사하는 방법이다 (Ahn 등, 2003). 3절에서 소월 (김정식)의 현대시들의 특성을 PPM으로 조사하는 과정과 유사한 것들을 찾는 방법을 논하고, 4절에서 설문지 조사 자료의 분석에 PPM을 적용하는 방법을 논한다.

* 이 논문은 2006년도 강릉대학교 교수연구년 연구지원에 의하여 수행되었음.

1) (210-702) 강원도 강릉시 지변동 123, 강릉대학교 정보통계학전공, 교수
E-mail: jsahn@kangnung.ac.kr

2. PP-자료와 그의 성질

상대도수로 표현된 2×2-분할표 $M_1=\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$로부터 다음과 같은 세 방향의 합을 생각하자.

$$\begin{pmatrix} a_1+b_1 \\ c_1+d_1 \end{pmatrix} \begin{pmatrix} a_1+c_1 \\ b_1+d_1 \end{pmatrix} \begin{pmatrix} a_1+d_1 \\ b_1+c_1 \end{pmatrix}.$$

위에서 처음 두 벡터는 주변 합이고 세 번째 벡터는 대각선 합이다. 각 벡터는 M_1의 원소가 오직 한번 나타나고 6개의 합들은 서로 다름을 볼 수 있다. 이들 합이 PP-값이고 첫 번째 원소들의 벡터 $(a_1+b_1\ a_1+c_1\ a_1+d_1)'$과 두 번째 원소들의 벡터 $(c_1+d_1\ b_1+d_1\ b_1+c_1)'$이 PP-자료 벡터 이다. $a_1+b_1+c_1+d_1=1$이므로 두 벡터 중 하나만 조사하여 분할표 자료의 특성을 비교할 수 있다. $\mathbf{q}_{11}=(a_1+b_1\ a_1+c_1\ a_1+d_1)'$, $\mathbf{q}_{12}=(c_1+d_1\ b_1+d_1\ b_1+c_1)'$라 하자. 같은 방법으로 $M_2=\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}$라 두면 $\mathbf{q}_{21}=(a_2+b_2\ a_2+c_2\ a_2+d_2)'$, $\mathbf{q}_{22}=(c_2+d_2\ b_2+d_2\ b_2+c_2)'$를 얻을 수 있다.

이때 $\mathbf{q}_{11}$과 $\mathbf{q}_{21}$의 거리제곱, $(a_1-a_2+b_1-b_2)^2+(a_1-a_2+c_1-c_2)^2+(a_1-a_2+d_1-d_2)^2$, 은 $(a_1\ b_1\ c_1\ d_1)'$과 $(a_2\ b_2\ c_2\ d_2)'$의 거리제곱, $(a_1-a_2)^2+(b_1-b_2)^2+(c_1-c_2)^2+(d_1-d_2)^2$과 같음을 볼 수 있다. 이 성질을 일반화하고 증명하기 위해 몇 가지 용어를 정의한다.

분할표에서 셀들을 원소로 갖는 집합을 라인 (line)이라 하고 두 라인이 공통원소를 갖지 않을 때 평행 (parallel)이라 한다. 평행인 라인들의 모든 원소가 분할표의 모든 셀과 같은 경우 이 라인들의 집합을 평행족 (parallel class)이라 한다. 단, 각 라인의 원소의 수는 모두 같다.

위의 M_1에서 $\{a_1\ b_1\}$과 $\{c_1\ d_1\}$은 평행이고 $\{\{a_1\ b_1\},\ \{c_1\ d_1\}\}$은 평행족이다. $\{a_1\ b_1\}$는 첫 번째 라인 (1st line)이고 $\{c_1\ d_1\}$는 두 번째 라인 (2nd line)이다.

두 평행족 C_1과 C_2에서 임의의 두 라인 $l_1\in C_1$, $l_2\in C_2$은 오직 하나의 원소를 공유한다 (Ahn 등, 2003, 참조). c^d-분할표에서 c가 소수일 때 이러한 평행족은 다음 함수로부터 $(c^d-1)/(c-1)$개 존재한다 (Laywine과 Mullen, 1998).

$$f_{\mathbf{a}}(\mathbf{x})=\mathbf{a}'\mathbf{x}=a_1x_1+\cdots+a_dx_d\equiv j(\mathrm{mod}\ \ c), \tag{2.1}$$

단, $\mathbf{a}$는 방향 벡터이고, $\mathbf{x}$은 범주형 변수 벡터이다.

어느 셀이 라인에 포함되면 1, 포함되지 않으면 0으로 둔 행렬을 P-행렬이라 한다. (2.1)식에서 $j=i$에 대한 P-행렬을 P^i라 하면 $\sum_{i=0}^{c-1}P^i=J_{\frac{c^d-1}{c-1}\times c^d}$이다. 여기서 J는 모든 원소가 1인 행렬이다. 이 때 i째 라인의 PP-자료 벡터는 $P^i\mathbf{x}$이다. 상대도수로 표현된 두 c^d-분할표 자료를 벡터 $\mathbf{r_1}$과 $\mathbf{r_2}$로 표현하면 다음 정리가 성립한다.

정리 2.1 $\sum_{i=0}^{c-1}[P^i(\mathbf{r_1}-\mathbf{r_2})]'[P^i(\mathbf{r_1}-\mathbf{r_2})]=c^{d-1}(\mathbf{r_1}-\mathbf{r_2})'(\mathbf{r_1}-\mathbf{r_2})$이다.

여기서 $P^i=(\mathbf{p}_1^i,\ \mathbf{p}_2^i,\ \cdots,\ \mathbf{p}_{c^d}^i)$, $\mathbf{p}_j^i$ 는 $(c^d-1)/(c-1)$개의 0 또는 1을 원소로 갖는 열벡터이다.

증명: $\sum_{i=0}^{c-1} P^{i\prime} = J_{c^d\times(c^d-1)/(c-1)}$ 임으로, $\sum_{i=0}^{c-1} P^{i\prime}P^i$의 대각원소는 $P^{i\prime}(i=1,2,\ldots,c-1)$의 열의 수 $(c^d-1)/(c-1)$와 같다. c 개의 $r\times n^d$행렬 $Q^{(j)} = (\mathbf{p}_1^i,\ \mathbf{p}_2^i,\ \ldots,\ \mathbf{p}_r^i)\prime$, $i = 0,1,\ldots,c-1$라 두자. 여기서 $\mathbf{p}_k^i$는 c^{d-1}개의 1과 $c^d - c^{d-1}$개의 0을 원소로 갖는 열벡터이고 $\sum_{i=0}^{c-1}\mathbf{p}_k^i = (1\ 1\ \ldots\ 1)\prime, \sum_{i=0}^{c-1} Q^i = J$ 이다. $Q^i = (\mathbf{q}_1^i,\ \mathbf{q}_2^i,\ \ldots,\ \mathbf{q}_{c^d}^i)$라 두고 Q^i의 j째 행벡터를 $(q_{j1}^i,\ q_{j2}^i,\ \ldots,\ q_{jc^d}^i)$라 하면 $i=0,1,\ldots,c-1$에 대한 c개의 행벡터는 각각 c^{d-1}개의 1을 갖는다. 이 때 모든 $k\neq l$에 대해 $\sum_{i=0}^{c-1}\mathbf{q}_k^{i\prime}\mathbf{q}_l^i = 0$되는 r의 최대값은 c^{d-1}이다. 따라서 $r=(c^d-1)/(c-1)$일 때 P-행렬의 성질 (Ahn 등, 2003, 참조)로부터 c가 소수일 때 모든$k\neq l$에 대해 $\sum_{i=0}^{c-1}\mathbf{q}_k^{i\prime}\mathbf{q}_l^i = (c^d-1)/(c-1) - c^{d-1} = (c^{d-1}-1)/(c-1)2$ 임을 알 수 있다. 고로 $\sum_{i=0}^{c-1}P^{i\prime}P^i$의 (k,l) 비대각원소는 $\sum_{i=0}^{c-1}\mathbf{p}_k^{i\prime}\mathbf{p}_l^i = (c^{d-1}-1)/(c-1)$ 이다. 한편 $(c^d-1)/(c-1) - (c^{d-1}-1)/(c-1) = c^{d-1}$ 이므로

$$\sum_{i=0}^{c-1} P^{i\prime}P^i = c^{d-1}I_{c^d\times c^d} + \frac{c^{d-1}-1}{c-1}\,J_{c^d\times c^d} \tag{2.2}$$

이다. 따라서 $\sum_{i=0}^{c-1}[P^i(\mathbf{r_1}-\mathbf{r_2})]\prime[P^i(\mathbf{r_1}-\mathbf{r_2})] = (\mathbf{r_1}-\mathbf{r_2})\prime[\sum_{i=0}^{c-1}P^{i\prime}P^i](\mathbf{r_1}-\mathbf{r_2}) = (\mathbf{r_1}-\mathbf{r_2})\prime[c^{d-1}I_{c^d\times c^d} + (c^{d-1}-1)/(c-1)\ J_{c^d\times c^d}](\mathbf{r_1}-\mathbf{r_2})$이고, $(\mathbf{r_1}-\mathbf{r_2})\prime J = (00\cdots0)_{1\times c^d}$이므로 $\sum_{i=0}^{c-1}[P^i(\mathbf{r_1}-\mathbf{r_2})]\prime[P^i(\mathbf{r_1}-\mathbf{r_2})] = c^{d-1}(\mathbf{r_1}-\mathbf{r_2})\prime(\mathbf{r_1}-\mathbf{r_2})$가 성립한다.

□

예를 들어 $c=3$, $d=2$인 경우 3개의 라인에 대한 P-행렬은 다음과 같다.

$$P^0 = \begin{pmatrix} 1&1&1&0&0&0&0&0&0\\ 1&0&0&1&0&0&1&0&0\\ 1&0&0&0&0&1&0&1&0\\ 1&0&0&0&1&0&0&0&1 \end{pmatrix},\ P^1 = \begin{pmatrix} 0&0&0&1&1&1&0&0&0\\ 0&1&0&0&1&0&0&1&0\\ 0&1&0&1&0&0&0&0&1\\ 0&1&0&0&0&1&1&0&0 \end{pmatrix},$$

$$P^2 = J - P^0 - P^1.$$

또한 (2.2)식은 다음과 같이 표현된다.

$$P^{0\prime}P^0 + P^{1\prime}P^1 + P^{2\prime}P^2 = \begin{pmatrix} 4&1&1&1&1&1&1&1&1\\ 1&4&1&1&1&1&1&1&1\\ 1&1&4&1&1&1&1&1&1\\ 1&1&1&4&1&1&1&1&1\\ 1&1&1&1&4&1&1&1&1\\ 1&1&1&1&1&4&1&1&1\\ 1&1&1&1&1&1&4&1&1\\ 1&1&1&1&1&1&1&4&1\\ 1&1&1&1&1&1&1&1&4 \end{pmatrix}.$$

3. 현대시의 사영탐색

현대시의 구조적 특성은 시의 형식 (민요시, 산문시), 표현 (은유적, 직설적), 율격구조

(6.4조, 7.5조), 음절의 길이, 작가의 시성 등이 복합적으로 나타난 결과로 볼 수 있다. 그러한 구조적 특성에 따라 독자에게 전달되는 느낌이 달라질 수 있을 것이다. 음율적 구조는 (정한모 등, 1982; 송하선, 1998)에서 국문학적인 관점에서 조사 연구되었다. 본 절에서 현대시로부터 분할표 자료를 생성하고 분할표 자료를 변환한 PP-자료의 플롯을 이용하여 시의 구조적 특성을 조사하는 과정을 설명한다. 그리고 12개의 소월 시에 적용할 것이다. 현대시의 PP-플롯은 시의 율격 구조와 작가가 선호한 단어의 음절 길이에 영향을 많이 받을 것이다. 또한 수형도와 분산이 가장 큰 두 방향에 대한 산점도를 PP-플롯의 유사성을 조사하는 보조 그래프로 사용할 수 있음을 보일 것이다.

3.1. 분할표 자료 생성

PP-플롯을 이용하여 현대시의 구조적 특성을 탐색하기 위해 먼저 분할표 자료를 생성해야한다. 분할표 자료 생성방법을 60음절이하인 소월 시를 사용하여 설명한다.

1단계) 주어진 시에서 각 음절의 글자 수를 조사한다.

2단계) 범주의 수 (c)를 정한다.

범주의 수와 범주를 나누는 경계 값을 정하는 객관적인 기준은 없고 분석자의 분석목적에 따라 달라질 수 있을 것이다. 본 절에서 인용한 12개의 소월 시는 표 3.1에서 보인 바와 같이 글자 수가 2보다 작거나 같은 음절의 평균비율이 0.49로 절반정도 됨으로 2보다 작거나 같으면 짧은 음절로 취급하여 범주 0을 주고 3과 같거나 크면 긴 음절로 취급하여 범주 1을 준다.

3단계) d개의 음절 쌍을 조사하여 c^d-분할표를 만든다.

시를 감상할 때 연 (또는 문장)의 구별이 중요한 요소이므로, 한 연 (또는 문장)에서 d음절씩 짝을 지우고 짝지어지지 않은 부분은 버리거나 적당한 방법으로 음절수를 추가하여 짝을 만든다. d음절의 순서쌍들을 d차원 분할표의 셀 $(x_1, x_2, \ldots, x_d)$의 객체로 보고 c^d-분할표를 만든다. d는 음절수의 개수를 고려하여 정한다.

위의 과정을 개벽 1922년 7월호에 발표된 소월 시 "진달래 꽃"에 적용하여 설명한다. 시집에 따라 띄어쓰기가 다르게 표현된 경우가 있음을 볼 수 있으나 본 절에서는 김정식 (1977)의 띄어쓰기를 참조한다. 다음 시 "진달래 꽃"의 1연, 3연, 4연은 율격이 7.5조를 이루고 있음을 볼 수 있다.

"나 보기가 역겨워 가실 때에는 말 없이 고이 보내 드리우리다.
영변에 약산 진달래꽃 아름 따다 가실 길에 뿌리우리다.
가시는 걸음 걸음 놓인 그 꽃을 사뿐히 즈려밟고 가시옵소서.
나 보기가 역겨워 가실 때에는 죽어도 아니 눈물 흘리우리다."

위의 시에서 연을 콤마로 구별한 음절의 글자 수는 다음과 같다.

1 3 3 2 3 1 2 2 2 5, 3 2 4 2 2 2 2 5, 3 2 2 2 1 2 3 4 5, 1 3 3 2 3 3 2 2 5

$c = 2$라 두고 글자 수가 2보다 작거나 같으면 범주 0에 대응시키고, 글자 수가 2보다 크면 범주 1을 대응시키면 다음과 같다.

0 1 1 0 1 0 0 0 0 1, 1 0 1 0 0 0 0 1, 1 0 0 0 0 0 1 1 1, 0 1 1 0 1 1 0 0 1

$d = 3$라 두고 다음과 같이 각 셀에 대응된 도수를 구할 수 있다. 단, 1연에서 3개씩 짝 지우고 남은 1개는 버리고 2연에서 남은 2개는 랜덤으로 정한 하나의 값 0을 추가하여 표를 만든 것이다.

셀	(1 1 1)	(1 1 0)	(1 0 1)	(1 0 0)	(0 1 1)	(0 1 0)	(0 0 1)	(0 0 0)
도수	1	0	1	1	3	2	1	3

3.2. 소월 시의 PP-플롯

본 절에서 소월 시 중에서 음절수가 60미만인 12개의 시 "1. 가는 길 (ganeun), 2. 가시나무 (gasi), 3. 개여울의 노래 (gae), 4. 고적한 날 (gojuk), 5. 먼 후일 (mern) , 6. 못 잊어 (mots), 7. 봄밤 (bombam), 8. 산유화 (san), 9. 예전엔 미처 몰랐어요 (yejun), 10. 왕십리 (wang), 11. 진달래 꽃 (jindal), 12. 합장 (habjang)"의 PP-플롯을 만들고 이들의 유사성을 조사한다.

위의 12개 시 들의 음절수와 범주 0에 속한 짧은 음절의 비율은 표 3.1과 같이 주어진다.

표 3.1: 음절수와 짧은 음절비율

시	1	2	3	4	5	6	7	8	9	10	11	12	평균
음절수	33	34	52	44	28	34	36	38	32	51	36	59	39.75
비 율	.52	.41	.44	.57	.32	.35	.42	.82	.50	.49	.56	.54	0.49

위의 12개 시 들의 2^3-분할표는 표 3.2와 같다. 표 3.2로부터 상대도수를 구하고 P-행렬을 사용하여 구한 PP-자료는 표 3.3로 주어진다. 표 3.3으로부터 PP-플롯은 그림 3.1에 주어지고 Ward 방법에 의한 수형도와 분산이 가장 큰 다섯 번째 방향 ($D5$)과 두 번째로 분산이 큰 네 번째 방향 ($D4$)의 산점도는 그림 3.2에 주어진다.

그림 3.1에서 자연을 노래하는 "산유화 (san)"와 그리움을 노래하는 "예전엔 미처 몰랐어요 (yejun)"는 넷째 방향까지 유사한 형태 (높이는 다름)를 이루고 있다. 이는 두 시 모두 4연시로 음율이 유사하기 때문인 것으로 생각된다. 그러나 그림 3.2의 수형도에서 "산유화 (san)"는 어느 시와도 집락을 이루지 않고 있으며 다른 11개의 시와 다른 구조적 특징을 갖고 있다.

간접묘사의 방법을 택하고 있는 "고적한 날 (gojuk)"과 직접묘사의 방법을 택하고 있는 "진달래 꽃 (jindal)"은 그림 3.1의 PP-플롯에서 높낮이 변화가 적음을 볼 수 있다 (정한모 등, 1982, 참조). 또한 "진달래 꽃"과 "봄밤", "가시나무"와 "먼 후일" 등의 PP-플롯 형태는 유사함을 볼 수 있다. 그림 3.2의 수형도에서 "산유화"를 제외한 11개 시는 크게 두 집락으로 나누어지는 것을 볼 수 있다. 한편 ($D5$, $D4$)의 산점도에서 두 집락의 구별은 분명하

표 3.2: 분할표

셀	1	2	3	4	5	6	7	8	9	10	11	12
1 1 1	0	3	3	2	3	2	1	0	1	5	1	4
1 1 0	2	2	3	1	3	3	2	1	0	1	0	2
1 0 1	0	0	0	0	0	1	2	0	3	0	1	3
1 0 0	1	1	3	1	0	1	1	3	2	1	1	0
0 1 1	3	1	2	1	1	2	3	1	0	1	3	0
0 1 0	2	0	2	3	1	0	2	0	3	3	2	4
0 0 1	3	1	4	2	1	3	0	0	1	3	1	1
0 0 0	1	2	0	4	1	0	1	9	1	2	3	6

표 3.3: PP-자료

방향	1	2	3	4	5	6	7	8	9	10	11	12
1	.25	.60	.53	.29	.60	.58	.50	.29	.55	.44	.25	.45
2	.58	.60	.59	.50	.80	.58	.67	.14	.36	.63	.50	.50
3	.50	.50	.53	.36	.50	.67	.50	.07	.45	.56	.50	.40
4	.50	.20	.41	.36	.20	.33	.67	.29	.73	.31	.58	.35
5	.75	.50	.71	.36	.50	.75	.50	.36	.27	.38	.42	.15
6	.58	.30	.53	.43	.50	.58	.50	.07	.64	.44	.33	.50
7	.50	.50	.71	.57	.50	.50	.33	.21	.64	.75	.42	.45

지 않으나 PP-플롯에서 유사한 것이 인접해 있음을 볼 수 있다. 그림 3.2는 그림 3.1을 해석하는데 유용한 보조 그래프이다.

음절수가 충분히 많은 경우 음절을 $c(>2)$가지로 구분하고 $c-1$개의 line에 대한 PP-플롯을 조사한다.

4. 설문자료의 사영탐색

4.1. 자료 생성

설문지의 문항을 변수로 지문을 범주로 취급하고 분할표를 만든다. 문항수가 d이고 지문수가 모두 c인 설문지 조사 자료로부터 c^d-분할표를 만들 수 있다. 예를 들어 초등학교 5, 6학년 학생 200명에게 보호자의 학력 (대졸, 고졸, 중졸, 초등졸)과 두 가지 지문 (가, 나)을 가진 4개의 문항 (1, 2, 3, 4)에 대해 답을 하게 한 결과가 다음과 같다고 하자 (자료출처: 통계 상담을 의뢰받았던 자료 중 4개의 문항을 발췌한 것임).

표 4.1에서 1행의 '22 38 12 10' 은 대졸, 고졸, 중졸, 초등졸인 보호자를 가진 학생 중 22명, 38명, 12명, 10명이 네 문항 모두 '가'에 답한 것을 나타낸다. 표 4.1은 '가'를 0으로 '

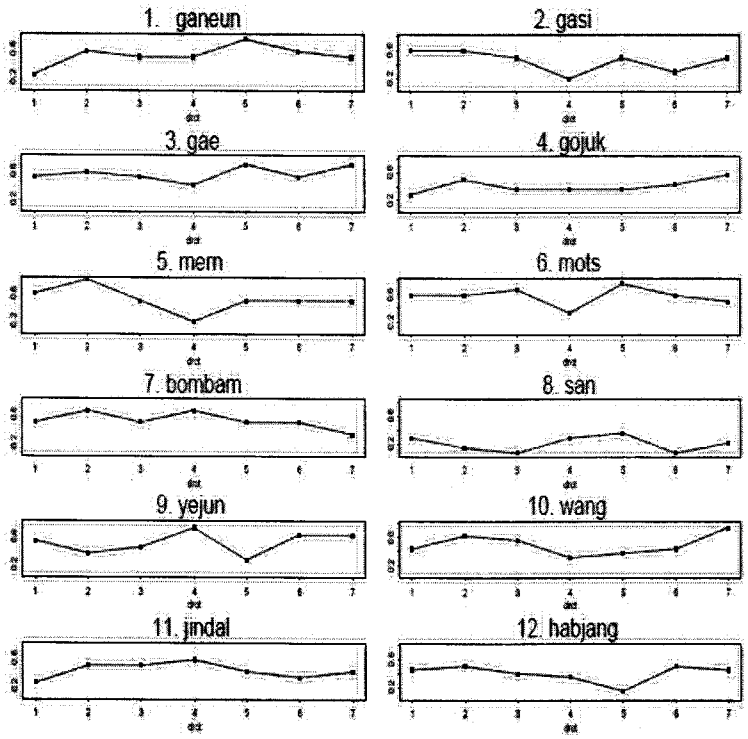

그림 3.1: 소월 시의 PP-플롯

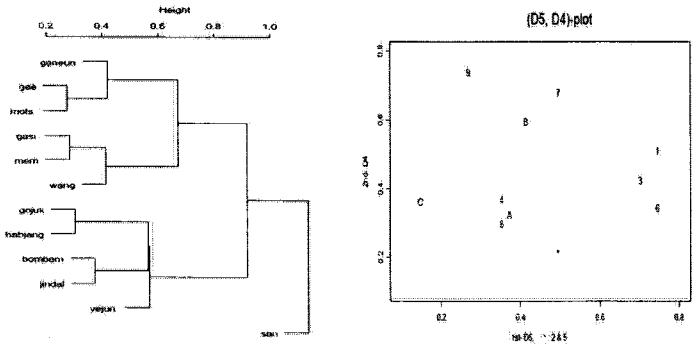

그림 3.2: 소월 시의 수형도와 $(D5, D4)$의 산점도

나'를 1로 바꾸면 0000셀에서 1111셀까지 16개의 셀을 가진 네 집단의 2^4-분할표이다. 이 2^4-분할표로부터 15개 방향에 대한 PP-값을 구하면 표 4.2와 같다.

표 4.1: 설문자료

문	항	번	호	학	력			문	항	번	호	학	력		
1	2	3	4	대	고	중	초	1	2	3	4	대	고	중	초
가	가	가	가	22	38	12	10	나	가	가	가	1	4	3	5
가	가	가	나	1	2	2	1	나	가	가	나	0	1	0	1
가	가	나	가	17	32	12	7	나	가	나	가	0	3	2	2
가	가	나	나	0	3	0	0	나	가	나	나	0	1	0	0
가	나	가	가	2	4	0	0	나	나	가	가	0	1	2	0
가	나	가	나	1	0	0	0	나	나	가	나	0	0	0	0
가	나	나	가	1	2	1	2	나	나	나	가	1	0	0	0
가	나	나	나	0	0	0	1	나	나	나	나	0	0	0	0

표 4.2: *PP*-자료

방향	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
대졸	.04	.11	.41	.04	.11	.41	.43	.09	.11	.46	.48	.11	.46	.43	.48
고졸	.11	.08	.45	.08	.16	.47	.48	.14	.15	.44	.48	.20	.46	.47	.47
중졸	.21	.09	.44	.06	.18	.53	.47	.26	.15	.50	.44	.24	.59	.53	.50
초등졸	.28	.10	.41	.10	.38	.55	.31	.31	.14	.45	.45	.34	.52	.41	.48

4.2. 설문자료의 *PP*-플롯

표 4.2의 *PP*-자료에 대한 *PP*-플롯은 그림 4.1에 주어지고 Ward 방법에 의한 수형도와 분산이 큰 두 방향 D5와 D8에 대한 산점도는 그림 4.2에 주어진다.

그림 4.1로부터 보호자의 학력이 대졸 (Univ)과 고졸 (High) 집단의 *PP*-플롯은 매우 유사하며 초등졸 (Prim)인 경우는 다른 세 집단과 다른 것을 볼 수 있다. 또한 패턴 (pattern) 변화, 즉 유사성의 순서가 대졸, 고졸, 중졸 (Middle), 초등졸인 것도 볼 수 있다. 그림 4.2의 수형도와 ($D5$, $D8$)의 산점도에서도 같은 결과를 볼 수 있으며 특히 산점도에서 초등졸 (P점)은 상대적으로 멀리 떨어져 있다. 문항에 따라 지문수가 다른 경우 각 그룹에서 몇 개의 c^d-분할표를 만들고 각각의 *PP*-플롯을 통합 연결한 플롯으로부터 자료를 탐색할 수 있다. 예를 들어 지문 3개인 문항 5개, 지문 4개인 문항 1개의 설문지의 경우 4개 지문 중 3개를 선택한 4개의 3^6 -분할표로부터 만든 *PP*-플롯을 통합 연결한 그래프에서 설문자료의 특성을 조사할 수 있다.

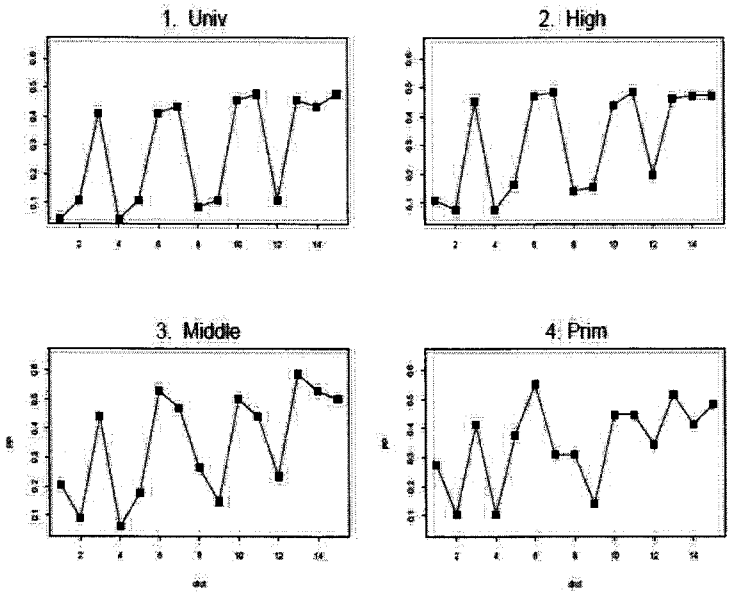

그림 4.1: 설문자료의 PP-플롯

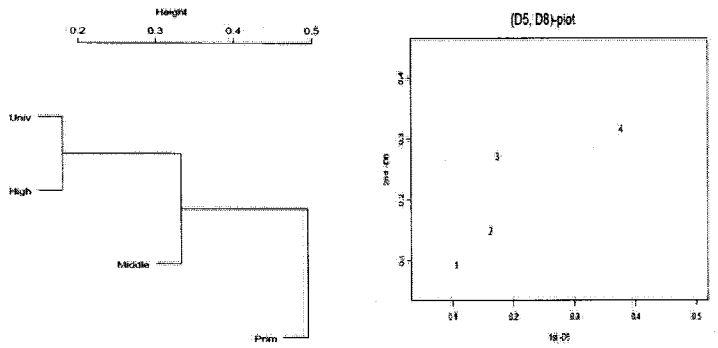

그림 4.2: 설문자료의 수형도와 $(D5, D8)$의 산점도

5. 결론 및 고찰

본 논문에서 Ahn 등 (2003)의 변환자료의 거리유지 성질을 보이고 $(c^d-1)/(c-1)$방향의 PP-값들로부터 그려진 PP-플롯을 현대시와 설문지 자료의 탐색에 응용하는 방법을 논했다. 또한 PP-플롯은 다차원 분할표 자료에서 거리보존의 특성을 갖고 플롯의 패턴으로 자료의 특성을 탐색할 수 있으므로 Andrews' curve의 이산형 다변량자료의 대조물(counterpart)로 볼 수 있다. 응용에서 소월 시 중 음절수가 적은 것을 $c=2$ $d=3$인 2^3-분할표 자료로 바꾸어 분석했다. 음절수가 충분히 많을 때 음절을 $c(>2)$가지로 구분하고 $c-1$개의 line에 대한 PP-플롯을 탐색하여 시들의 유사성과 특성을 조사할 수 있다. 현대시를 분할표 자료로 바꿀 때 가장 중요한 요소가 띄어쓰기와 연 (또는 문장)의 구별이다. 그런데 인용한 시집마다 이들이 약간씩 다르게 표현된 것을 볼 수 있다. 따라서 어느 책을 기준으로 했는가에 따라 약간의 차이가 있을 수 있으며 특히 짧은 시에서는 심각한 차이를 보일 수도 있다. 또한 c와 d의 선택에 따라 집락이 달라질 수 있기 때문에 적적할 값을 찾는 방법을 고려해야한다. 한편 설문자료의 탐색에서 응답자 수가 네 그룹에서 200명뿐임으로 $c=2$, $d=4$인 2^4-분할표 자료로 바꾸고 PP-플롯을 만들었다. 설문지의 지문 수와 문항 수가 큰 경우 조사자 수는 충분히 많아야 될 것이다. 또한 지문 수가 다른 경우도 몇 개의 PP-플롯을 통합 연결한 그래프로부터 설문조사 자료를 탐색할 수 있다.

참고문헌

김정식 (1977). <소월 시 전집>, 성공문화사.
송하선 (1998). <한국명시 해설>, 국학자료원.
정한모, 김열규, 신동욱 (1982). <김소월 연구>, 새문사.
Ahn, J. S., Hofmann, H. and Cook, D. (2003). A projection pursuit method on the multi-dimensional squared contingency table, *Computational Statistics*, **18**, 605–626.
Andrews, D. F. (1972). Plots of high-dimensional data, *Biometrics*, **28**, 125–136.
Laywine, C. F. and Mullen, G. L. (1998). *Discrete Mathematics using Latin Squares*, John Wiley & Sons, New York.

[2007년 6월 접수, 2007년 8월 채택]

The Transform of Multidimensional Categorical Data and its Applications*

Ju Sun Ahn[1)]

ABSTRACT

The squared Euclid distance of the values which is transformed by P-matrix of Ahn *et al.* (2003) is in proportion to the squared Euclid distance of cell's relative frequencies in two Contingency Tables. We propose the method of using the PP-values for the analysis of modern poems and questionnaire data.

Keywords: P-matrix, PP-plot, modern poem, questionnaire data.

* This work was supported by Kangnung National University Reserch Grant in 2006.

1) Professor, Department of Statistics, Kangnung National University, Korea
E-mail: jsahn@kangnung.ac.kr