응용통계연구 제 14권 2호, 2001년, pp. 393-399

행렬도를 이용한 개체와 변수간의 밀접도에 대한 연구

유성모[1] 김상우[2] 최강호[3]

요 약

자료행렬에서의 개체와 변수간의 관계 또는 분할표 자료에서의 열 범주와 행 범주간의 밀접도를 표준화된 자료행렬에 대한 요인행렬도에서 개체(행)와 변수(열)에 해당하는 두 벡터의 사이각의 코사인으로 정의하였다. 본 논문에서 정의한 개체와 변수간의 밀접도를 15대 및 16대 국회의원 선거자료에 적용하여 보았다.

주요용어: 밀접도, 지역성, 행렬도.

1. 서론

여러 변수에 대한 여러 개체의 반응치 또는 관찰치로 이루어진 자료행렬과 행범주와 열범주에 대한 빈도수로 이루어진 분할표 형태의 자료행렬에서 개체(또는 행범주)와 변수(또는 열범주)와의 관계성을 시각적으로 표현하기 위한 탐색적인 분석기법중 하나가 행렬도이다. 본 논문에서는 행렬도에서 시각적으로 표현되는 개체와 변수간의 관계의 밀접성을 수치적으로 나타내기 위한 밀접도를 정의하였다. 정의된 밀접도의 사례 적용을 위하여 15대 국회의원 선거(1996년 4월 11일) 및 16대 국회의원 선거(2000년 4월 13일) 자료에서 주요 정당의 지역별 득표수에 대한 분할표 자료행렬을 토대로 행렬도를 작성하였고, 지역과 주요 정당과의 밀접도를 구하였다. 행렬도의 작성과 밀접도의 계산은 S-PLUS 4.5를 이용하였다.

2. 개체와 변수간의 밀접도에 대한 정의

2.1. 행렬도

Gabriel(1971)에 의해서 주창된 행렬도(Biplot)에서는 자료행렬의 개체(행)와 변수(열)의 관계를 저차원(일반적으로 2차원)상의 그림으로 나타낸다. 자료행렬 $X_{n\times p}$는 비정칙치 분해에 의해서 계수가 r인 두 행렬, $G_{n\times p}$와 $H'_{p\times p}$의 곱으로 표현되며 이는 $G_{n\times 2}$와 $H'_{2\times p}$의 곱으로 근사된다. 즉,

1) (339-700) 충남 연기군 조치원읍, 고려대학교 정보통계학과, 부교수
E-mail: syoo@tiger.korea.ac.kr
2) (135-280) 서울시 강남구 대치동 946-18 대원빌딩 12층, (주)아이클릭 연구본부, 연구원
E-mail: miner@eyeclick.co.kr
3) (339-700) 충남 연기군 조치원읍, 고려대학교 정보통계학과, 석사과정
E-mail: random@korea.ac.kr

$$X_{n\times p} = U_{n\times p} D^{\alpha}{}_{p\times p} D^{1-\alpha}{}_{p\times p} V'_{p\times p} = G_{n\times p} H'_{p\times p} \simeq G_{n\times 2} H'_{2\times p}$$

이다. 여기서, U와 V는 각각 $U'U = I_p$와 $V'V = VV' = I_p$ 를 만족하는 직교행렬이고 D는 자료행렬 $X_{n\times p}$의 공분산 행렬 $S = \frac{1}{n-1}(X-\underline{\mu})'(X-\underline{\mu})$에 대한 고유치의 제곱근들을 원소로 하는 대각행렬이다. 또한, $G = UD^{\alpha}$이고 $H' = D^{1-\alpha}V'$이며, α는 0 또는 1이다. α 값에 따라 G와 H'의 형태는 달라진다. $\alpha = 0$ 인 경우를 요인행렬도라 하며, $\alpha = 1$ 인 경우를 주성분행렬도라 한다. 중심화된 자료행렬을 이용할 경우 요인행렬도의 G행렬을 통해서는 개체간의 마할라노비스(Mahalanobis) 거리를 구할 수 있으며 H행렬을 통해서는 변수간의 상관관계를 파악할 수 있다. 주성분행렬도의 G행렬을 통해서는 개체간의 유클리드(Euclid)거리를 파악할 수 있다(허명회 1993, v-8 참조).

2.2. 개체와 변수간의 밀접도

이미 주지하는 바와 같이 $\alpha = 0$ 인 경우의 요인행렬도에서는 중심화된 자료행렬을 이용할 경우 G행렬을 통해서는 개체간의 마할라노비스 거리를 구할 수 있으며 H행렬을 통해서는 변수간의 상관관계를 파악할 수 있고, $\alpha = 1$ 인 경우의 주성분행렬도에서는 자료행렬을 그대로 이용할 경우 G행렬을 통해서는 개체간의 유클리드 거리를 파악할 수 있다. 또한 표준화된 자료행렬을 이용할 경우 마할라노비스 거리와 유클리드 거리는 같다. 이러한 사실에 착안하여 표준화된 자료행렬에 대한 요인행렬도에서 i번째 개체(행)와 j번째 변수(열)에 해당하는 표준화된 자료 x^s_{ij}를 구성하는 두 벡터의 사이각의 코사인을 i번째 개체(행)와 j번째 변수(열)의 밀접도 r_{ij}로 정의하고자 한다. 즉,

$$r_{ij} = \cos\theta_{ij} = \frac{\vec{g}_i \cdot \vec{h}'_j}{|\vec{g}_i||\vec{h}'_j|}$$

이다. 여기서,

$$\begin{aligned} x^s_{ij} &\simeq g_{i1}h_{j1} + g_{i2}h_{j2} \\ &= \vec{g}_i \cdot \vec{h}'_j = |\vec{g}_i||\vec{h}'_j|\cos\theta_{ij} \\ &\quad i = 1,\cdots,n;\ \ j = 1,\cdots,p;\ \ 0^\circ \le \theta_{ij} \le 180^\circ \end{aligned}$$

이다. 위에서 정의한 밀접도는 -1과 +1 사이의 값을 가지며, 밀접도 r_{ij}가 +1에 가까울수록 i번째 개체(행)와 j번째 변수(열)는 서로 밀접한 양의 관계가 있다는 것을 의미하며, 밀접도 r_{ij}가 -1에 가까울수록 i번째 개체(행)와 j번째 변수(열)는 서로 밀접한 음의 관계가 있다는 것을 의미하고, 밀접도 r_{ij}가 0에 가까울수록 i번째 개체(행)와 j번째 변수(열)는 서로의 관계가 미약하다는 것을 의미한다.

본 논문에서 정의한 밀접도는 개념적으로는 임의의 두 벡터가 이루는 사이각의 코사인이다. 또한, 두 변수 사이의 상관계수는 두 변수 벡터가 이루는 사이각의 코사인이라는 것은 주지의 사실이다. 따라서, 일반적인 상관계수는 본 논문에서 정의한 밀접도의 한 예라고 볼 수 있다.

2.3. 적용사례

15대 및 16대 국회의원 선거에서 각 정당의 지역별 득표수를 토대로 한국 정당의 지역성에 대한 문제를 고찰해 보기로 하자. 각 정당의 15대 및 16대 국회의원 선거에서의 지역별 득표수는 분할표 형태의 자료이다. 특정 정당의 득표수는 정당에 따라서 차이가 있기 때문에, 특정 정당이 특정 지역에서 지지를 받는 정도는 특정 정당의 득표수에 따라서 표준화되어야 한다. 이러한 표준화의 한 방법으로는 특정 정당에 대한 전체 득표수를 토대로 특정 정당의 지역별 득표율을 구하는 방법과 정당과 지역간의 독립성을 토대로 한 특정 정당의 특정 지역에서의 표준화된 개체 값을 이용하는 방법이 있다. 본 논문에서는 표준화된 개체 값(χ_{ij})을 이용하였으며, 이의 공식은 다음과 같다.

$$\chi_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \frac{n_{ij} - n_{i.}n_{.j}/n_{..}}{\sqrt{n_{i.}n_{.j}/n_{..}}}, \quad i = 1, \cdots, n; \quad j = 1, \cdots, p$$

여기서 O_{ij}는 분할표에서의 실제 개체이고, E_{ij}는 지역과 정당간의 관계가 없다는 가설하에서의 기대치를 말한다. 위에서 정의한 표준화된 관측치 χ_{ij}는 대응분석에서의 대응 행렬의 원소 m_{ij}에 상수 $\sqrt{n_{..}}$을 곱한 것과 같다. 즉, $\chi_{ij} = m_{ij}\sqrt{n_{..}}$이다. 본 논문에서 정의한 표준화된 관측치의 값 χ_{ij}는 정당과 지역 간에 관계가 없다는 가정에서 i번째 개체의 j번째 변수에 해당되는 값이 상대적으로 기대되는 값에서 벗어난 정도를 나타내는 값으로 해석할 수 있으며, χ_{ij}의 값이 같은 개체의 다른 어떤 변수에 해당되는 값보다 크거나 같은 경우(즉, $\chi_{ij} \geq \chi_{ij'}, j' \neq j = 1, \cdots, p$) i번째 개체는 j번째 변수와 양의 방향으로 밀접한 관계에 있다고 볼 수 있다. 본 논문에서는 표준화된 행렬을 정당별로 구한 평균과 표준편차를 이용하여 다시 표준화한 행렬에 대한 비정칙치 분해를 토대로 요인행렬도를 작성하였다. 표준화된 자료행렬을 이용할 경우에 요인행렬도의 G행렬을 통해서는 개체간의 마할라노비스 거리를 구할 수 있기 때문에 지역 간의 유사성을 나타내는 척도로 사용될 수 있으며, H행렬을 통해서는 변수간의 상관관계를 파악할 수 있기 때문에 정당간의 상관관계를 파악할 수 있다. 지역성을 나타내는 G행렬 벡터는 행렬도의 좌표 위에 지역을 나타내는 레이블로 표시하였고 정당을 나타내는 H행렬 벡터는 원점과 좌표를 연결하는 직선으로 표시하였다. 또한 표준화된 관측치 값을 구할 때에는 무소속에 대한 득표수를 고려하였지만 주요 정당과 지역과의 밀접성을 알아보기 위한 행렬도의 작성에서는 무소속을 제외한 주요 정당의 자료만을 이용하였다. 15대, 16대 국회의원 선거결과를 이용하여 표준화된 관측치의 값 χ_{ij}에 대한 요인행렬도를 작성하면 그림 2.1과 그림 2.2와 같다.

이미 언급했듯이 그림 2.1과 그림 2.2에서의 각 지역 간의 거리는 지역 간의 마할라노비스 거리를 나타낸다. 마할라노비스 거리가 가까울수록 지역 간의 유사성이 높음을 나타내고 마할라노비스 거리가 멀수록 지역 간의 유사성은 낮음을 의미한다. 특정 지역들간에는 유사성이 높음을 알 수 있고, 특정 지역과 특정 정당간의 밀접성은 두 벡터를 이용하여 정의한 밀접도를 통하여 알 수 있다. 15대 및 16대 국회의원 선거자료에 대하여 지역과 정당간의 밀접도를 구하면 표 2.1과 같다.

국회의원 선거에서의 지역성이란 특정 정당이 특정 지역에 밀접성이 매우 큰 경우를 의미한다. 이는 본 논문에서 정의한 밀접도 r_{ij}의 값이 +1에 가까울 수록 j번째 정당이 i번

표 2.1: 15대, 16대 국회의원 선거의 지역과 정당간의 밀접도

15대	신한국당 (SHINHAN)	국민회의 (KUKMIN)	민주당 (MINJU)	자민련 (JAMIN)	16대	한나라당 (HANNARA)	새천년민주당 (SAEMIN)	자민련 (JAMIN)	민주국민당 (MINKUK)
서울 (SL)	0.492	0.341	0.309	-0.999	서울 (SL)	-0.254	0.633	-0.889	-0.386
부산 (PS)	0.986	-0.517	0.934	-0.608	부산 (PS)	0.967	-0.773	-0.460	0.921
대구 (TG)	-0.540	-0.287	-0.362	0.997	대구 (TG)	0.974	-0.980	0.010	0.996
인천 (IC)	0.574	0.248	0.400	-0.992	인천 (IC)	-0.975	0.794	0.430	-0.934
광주 (GJ)	-0.587	0.997	-0.736	-0.440	광주 (GJ)	-0.769	0.965	-0.457	-0.851
대전 (DJ)	-0.345	-0.488	-0.152	0.991	대전 (DJ)	-0.362	-0.058	0.989	-0.228
경기 (KK)	0.998	-0.694	0.990	-0.417	경기 (KK)	-0.937	0.707	0.545	-0.879
강원 (KW)	0.705	-0.997	0.832	0.295	강원 (KW)	0.998	-0.883	-0.274	0.980
충북 (CB)	-0.244	-0.578	-0.047	0.971	충북 (CB)	-0.719	0.365	0.835	-0.615
충남 (CN)	-0.118	-0.677	0.081	0.932	충남 (CN)	-0.261	-0.164	0.999	-0.124
전북 (JB)	-0.785	0.981	-0.892	-0.178	전북 (JB)	-0.762	0.962	-0.467	0.845
전남 (JN)	-0.864	0.945	-0.947	-0.039	전남 (JN)	-0.800	0.977	-0.412	-0.876
경북 (KB)	0.054	-0.793	0.251	0.856	경북 (KB)	.865	-0.995	0.302	0.927
경남 (KN)	0.969	-0.442	0.900	-0.674	경남 (KN)	0.961	-0.760	-0.478	0.913
제주 (JJ)	0.026	0.742	-0.173	-0.895	제주 (JJ)	-0.221	0.607	-0.904	-0.356
울산 (UL)	-	-	-	-	울산 (UL)	0.946	-0.725	-0.522	0.891

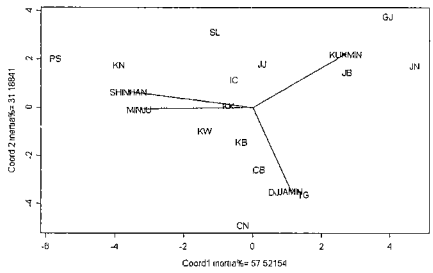

그림 2.1: 15대 국회의원 선거결과에 대한 요인행렬도

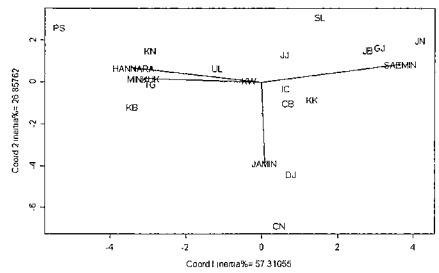

그림 2.2: 16대 국회의원 선거결과에 대한 요인행렬도

째 지역에서 상대적으로 강력한 지지를 받고 있다는 것을 의미하며, r_{ij}의 값이 -1에 가까울 수록 j번째 정당이 i번째 지역에서 상대적으로 저조한 지지를 받고 있다는 것을 의미한다. 표 2.1을 좀더 살펴보면 15대 국회의원 선거의 경우 신한국당의 지역별 밀접도는 경기(0.998), 부산(0.986), 경남(0.969), 강원(0.705)에서 +1에 가까운 값을 나타내었고, 전남(-0.864), 전북(-0.785)에서 -1에 가까운 값을 나타내었으며, 제주(0.026)에서 0에 가까운 값을 나타내었다. 이는 신한국당이 경기, 부산, 경남, 강원 지역에서 타 정당에 비해서 상대적으로 강력한 지지를 받고 있는 반면에, 전남, 전북지역에서 상대적으로 매우 저조한 지지를 받고 있으며, 제주지역에서는 중도적인 지지를 받고 있다는 것을 의미한다. 국민회의의 지역별 밀접도는 광주(0.997), 전북(0.981), 전남(0.945)에서 +1에 가까운 값을 나타내었고, 강원(-0.997), 경북(-0.793), 경기(-0.694), 충북(-0.578)에서 -1에 가까운 값을 나타내었다. 이는 국민회의가 광주, 전북, 전남지역에서 타 정당에 비해서 상대적으로 강력한 지지를 받고 있는 반면에, 강원, 경북, 경기, 충북지역에서 상대적으로 매우 저조한 지지를 받고 있음을 의미한다. 자민련의 지역별 밀접도는 대구(0.997), 대전(0.991), 충북(0.971), 충남(0.932), 경북(0.856)에서 +1에 가까운 값을 나타내었고, 서울(-0.999), 인천(-0.992), 제주(-0.895)에서 -1에 가까운 값을 나타내었다. 이는 자민련이 대구, 대전, 충북, 충남지역에서 타 정당에 비해서 상대적으로 강력한 지지를 받고 있는 반면에, 서울, 인천, 제주 지역에서 상대적으로 매우 저조한 지지를 받고 있음을 의미한다.

16대 국회의원 선거의 경우 한나라당의 지역별 밀접도는 강원(0.998), 대구(0.974), 부산(0.967), 경남(0.961), 울산(0.946)에서 +1에 가까운 값을 나타내었고, 인천(-0.975), 경기(-0.937), 전남(-0.800), 광주(-0.769), 전북(-0.762)에서 -1에 가까운 값을 나타내었으며, 밀접도에 대한 절대값의 크기로 볼 때, 서울(-0.254)에서 가장 작은 값을 나타내었다. 이는 한나라당이 강원, 대구, 부산, 경남, 울산지역에서 타 정당에 비해서 상대적으로 강력한 지지를 받고 있는 반면에, 인천, 경기, 전남, 광주, 전북 지역에서 상대적으로 매우 저조한 지지를 받고 있으며, 서울 지역에서 중도적인 지지를 받고 있다는 것을 의미한다. 새천년민주당의 지역별 밀접도는 전남(0.977), 광주(0.965), 전북(0.962), 인천(0.794)에서 +1에 가까운 값을 나타내었고, 경북(-0.995), 대구(-0.980), 부산(-0.773), 경남(-0.760)에서 -1에 가까운 값을 나타내었으며, 대전(-0.058) 및 충남(-0.164)에서 밀접도의 절대값의 크기로 볼 때 가장 작은

값을 나타내었다. 이는 새천년민주당이 전남, 광주, 전북, 인천지역에서 타 정당에 비해서 상대적으로 강력한 지지를 받고 있는 반면에, 경북, 대구, 부산, 경남지역에서 상대적으로 매우 저조한 지지를 받고 있으며, 대전, 충남지역에서는 중도적인 지지를 받고 있다는 것을 의미한다. 자민련의 지역별 밀접도는 충남(0.999), 대전(0.989), 충북(0.835)에서 +1에 가까운 값을 나타내었고, 제주(-0.904), 서울(-0.889)에서 -1에 가까운 값을 나타내었으며, 대구(0.010)에서 0에 가까운 값을 나타내었다. 이는 자민련이 충남, 대전, 충북지역에서 타 정당에 비해서 상대적으로 강력한 지지를 받고 있는 반면에, 제주, 서울지역에서 상대적으로 매우 저조한 지지를 받고 있으며, 대구 지역에서는 중도적인 지지를 받고 있다는 것을 의미한다.

본 적용 사례에서 구한 밀접도는 특정 정당의 특정 지역에 대한 상대적인 지역별 밀접도를 의미하기 때문에 특정 지역에 대한 두 정당의 밀접도들이 동일 할 지라도 그 두 정당의 특정 지역에 대한 득표수가 동일하다는 의미는 아니다. 따라서, 특정정당의 지역별 밀접도에 대한 해석에 신중을 기할 필요가 있다. 아울러 양대 국회의원 선거에서 나타난 각 정당과 지역 간의 밀접도에 대한 심도 깊은 해석은 정치학자들의 몫으로 남겨둔다.

3. 결론

본 논문에서는 행렬도를 이용하여 개체(행)와 변수(열)간의 밀접도를 정의하였다. 밀접도는 표준화된 자료행렬에 대한 요인행렬도에서의 개체(행)와 변수(열)의 표준화된 자료를 구성하는 두 벡터의 사이각의 코사인으로 정의하였다. 본 논문에서 정의한 밀접도의 예시를 위하여 양대 국회의원 선거자료를 이용하였다.

참고문헌

[1] 최용석 (1993). <SAS 대응분석>, 자유아카데미, 서울.

[2] 허명회 (1993). <통계상담의 이해>, 자유아카데미, 서울.

[3] 허명회 (1999). <다변량 수량화>, 자유아카데미, 서울.

[4] Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 3, p.453.

[5] MathSoft (1997). S-PLUS 4 Guide to Statistics.

[2001년 4월 접수, 2001년 7월 채택]

On the Closeness between an Observation and a Variable in a Biplot

Seongmo Yoo[1] Sang Woo Kim[2] Kang Ho, Choi[3]

ABSTRACT

The closeness representing the relationship between a observation and a variable in a data matrix or the relationship between row and column categories in a frequency table is defined as the cosine of the angle between a observation vector and a variable vector in a biplot. The results of the last two elections for members to Korean national assembly were analysed and interpreted for the application of the idea of closeness between a observation and a variable in the biplot. The regional closeness for the political parties is revealed and found to be intensified.

Keywords: Biplot; Closeness.

1) Associate Professor, Dept. of Informational Statistics, Korea University.
E-mail: syoo@tiger.korea.ac.kr
2) Statistician, EyeClick Inc.
E-mail: miner@eyeclick.co.kr
3) Graduate Student, Dept. of Informational Statistics, Korea University.
E-mail: random@korea.ac.kr