응용통계연구 제 **14**권 2호, *2001*년, *pp. 305–320*

다차원선호분석의 최적척도화 및 부분수량화*

황선영[1] 정수진[2] 김영원[3]

요 약

다차원선호분석(mutidimensional preference analysis)은 여러 상품들에 대한 개인(또는 그룹)의 선호도를 알아보기 위한 분석방법으로 결과는 보통 2차원 그림으로 제공된다. 본 연구에서는 의미 있는 두 가지 최적척도 기준을 제안하고 이와 연관된 행 및 열 표시자를 유도하고 있으며, 아울러 사전지식을 반영하기 위해 부분수량화를 다차원선호분석에 도입하는 방법을 제시한다. 또한 본 연구에서 제시한 다차원분석기법들을 실제 인터넷 검색엔진에 대한 선호도 자료에 적용한다.

주요용어: 다차원선호분석, 부분수량화, 최적척도화, 행렬도.

1. 서론

다차원선호분석(MDPREF: MultiDimensional PREFerence analysis)은 행렬도(biplot)의 일종으로서(Gabriel, 1971), 여러 상품들과 이들 상품에 대한 개인(또는 그룹)의 선호도를 알아보기 위한 분석 방법으로 결과는 보통 2차원 그림으로 제공된다. 다차원선호분석은 선호도 자료를 중심으로 한 분석기법으로, 이 선호도 자료는 소비자들의 상품에 대한 선호 혹은 상품 속성에 대한 각 상품별 평가를 말하는데 다차원선호분석은 이러한 두 종류의 대상(예를 들어, 소비자와 상품 혹은 속성과 상품)을 동시에 공통공간에 기하학적인 구조로 나타내 준다(임종원, 1997). 선호 데이터는 행과 열로 구성되어 있으며 일반적으로 행은 "상품"을 그리고 열은 "소비자, 소비자 그룹 또는 상품의 속성"을 나타낸다. 이러한 선호 데이터를 공통공간에 동시에 나타 낼 때 행(상품)은 점으로, 열은(소비자, 소비자 그룹 또는 상품의 속성) 벡터로 나타낸다(Kuhfeld, 1993). 본 논문에서는 이러한 다차원선호분석을 두 가지 측면에서 살펴보고자 한다.

첫째, 두 가지 최적척도 기준을 제안하고 이 기준하에서 최적(optimal)의 행표시자와 열표시자를 유도하고자 한다.

둘째, 연구자가 소비자들의 상품 선택에 있어서 가장 중요시 여기는 속성을 사전에 알고 있을 경우, 이를 제 1축으로 전제하고 나머지 축은 자료를 통해서 찾아내는 다차원선호분석의 부분수량화(partial quantification)를 수행하고자 한다.

* 본 연구는 KISTEP연구비(2000) 지원에 의해 수행되었음.

1) (140-742) 서울시 용산구 청파동, 숙명여자대학교 통계학과, 부교수
E-mail: shwang@sookmyung.ac.kr
2) (140-742) 서울시 용산구 청파동, 숙명여자대학교 통계학과, 대학원
3) (140-742) 서울시 용산구 청파동, 숙명여자대학교 통계학과, 교수
E-mail: ywkim@sookmyung.ac.kr

본 논문의 사례는 인터넷상에서 온라인으로 얻은 인터넷 검색엔진들에 대한 선호도 자료를 사용하였다. 매일경제 여론조사(1999년 9월), 마이넷 인터넷 설문조사(1998년 12월 15일 ~ 1999 년 1월 20일) 그리고 kissnet (1998년 4월 7일)의 설문조사 결과를 참고로 하여 가장 많이 사용되는 검색엔진인 까치네, 네이버, 라이코스 코리아, 심마니, 야후 코리아, 한미르, 한글 알타비스타를 조사대상으로 삼았다. 그리고 이러한 검색엔진 선택시 중요시 여기는 속성으로는 1997년 8월 26일 NPD Online Research(NPD Group Inc.)가 22,000명을 대상으로 조사한 온라인 설문조사 결과를 바탕으로 속도, 검색결과의 정확성, 사용방법의 용이성, 최신정보 보유량의 4가지를 사용하였다. 설문조사는 kissnet이라는 인터넷 리서치 전문회사에 의뢰하여, kissnet 홈페이지(http://www.kissnet.co.kr)에서 1999년 9월 17일에서 20일까지 4일에 걸쳐 총 200명에게 이루어졌다.

또한 검색엔진에 대한 선호도는 연령별로 차이가 있을 것으로 예상하고 전체 모집단을 초/중/고등학생, 대학(원)생, 30대 이전의 직장인, 30대 이후의 직장인의 4그룹으로 층화하였으며(본 논문에서는 이를 각각 그룹1, 2, 3, 4로 표현한다.), 각 층별로는 50명씩 동수로 추출하였다. 이렇게 얻은 총 200명의 자료 중 검색엔진별로 선호도의 차이가 없다고 응답한 33명의 데이터를 제외하고, 총 167명의 자료를 사용하여 분석하였다. 그룹별로 분석에 사용된 표본의 수는 초/중/고등학생 46명, 대학(원)생 40명, 30대 이전의 직장인 43명, 30대 이후의 직장인 38명이다.

2. 계량형 다차원 선호분석모형

본 연구에서는 선호자료가 구간척도를 만족한다고 가정하는 계량형(metric) 분석을 시도하고자 하며 이에 따라 먼저 계량형 다차원 선호분석모형에 관한 수식을 정리하고자 한다. 다차원 척도 문제에서 계량형과 비계량형(nonmetric)의 구분 및 연관된 알고리즘은 최용석(1995)를 참고하기 바란다.

n개의 상품(행)에 대한 m개 그룹(열)의 선호도 행렬을 $X_{n\times m} = \{X_{ij}\}$ $(i = 1, \cdots, n \; j = 1, \cdots, m)$ 이라 하고 행렬의 계수(rank)는 $r(r \le min(n, m))$ 로 표기하자. 그룹별 응답을 표준화하기 위해 입력자료를 각 열별로 먼저 표준화시키고,

$$z_{ij} = \frac{x_{ij} - m_j}{s_j} \tag{2.1}$$

여기서, m_j와 s_j는 각각 j번째 그룹의 평균 및 표준편차이다.

행과 열을 동시에 저차원으로 표현하기 위해 표준화된 자료행렬 $Z = \{z_{ij}\}$를 다음과 같이 비정칙값분해(singular value decomposition)하자.

$$Z_{n\times m} = P_{n\times r} \, \Delta_{r\times r} \, Q^T_{r\times m} = GH^T \tag{2.2}$$

여기서 $G = P\Delta^c$, $H = Q\,\Delta^{1-c}$ 이고, $0 \le c \le 1$ 이다. (2.2)에서 r은 행렬 Z의 계수이며, 행렬 $P = (\, p_1, \, p_2, \, \cdots, \, p_r \,)$ 의 열벡터는 ZZ^T의 정칙고유벡터(orthonormal eigenvector)들이고, 행렬 $Q = (\, q_1, \, q_2, \, \cdots, \, q_r \,)$ 의 열벡터는 Z^TZ 의 정칙고유벡터들이다. 즉, $P^TP = Q^TQ = I_{r\times r}$ 이다. 또한 $\Delta_{r\times r}$ 은 주대각원소가 크기 순으로 된 비정칙

값(singular values), $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r \geq 0$ 들을 갖는 주대각 행렬이며, 이들 비정칙 값들은 정방행렬 Z^TZ 또는 ZZ^T의 고유값들의 양의 제곱근이다.

(2.2)의 2차원 근사적 표현으로서 다음 식을 생각해 보자.

$$Z_{(2)} = G_{(2)} H_{(2)}^T \tag{2.3}$$

여기서

$$G_{(2)} = P_{(2)} \Delta_{(2)}^c = \begin{pmatrix} p_{11}\,\delta_1^c & p_{12}\,\delta_2^c \\ \vdots & \vdots \\ p_{i1}\,\delta_1^c & p_{i2}\,\delta_2^c \\ \vdots & \vdots \\ p_{n1}\,\delta_1^c & p_{n2}\,\delta_2^c \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} \\ \vdots & \vdots \\ g_{i1} & g_{i2} \\ \vdots & \vdots \\ g_{n1} & g_{n2} \end{pmatrix} = \begin{pmatrix} g_1^T \\ \vdots \\ g_i^T \\ \vdots \\ g_n^T \end{pmatrix}$$

$$H_{(2)} = Q_{(2)} \Delta_{(2)}^{1-c} = \begin{pmatrix} q_{11}\,\delta_1^{1-c} & q_{12}\,\delta_2^{1-c} \\ \vdots & \vdots \\ q_{j1}\,\delta_1^{1-c} & q_{j2}\,\delta_2^{1-c} \\ \vdots & \vdots \\ q_{m1}\,\delta_1^{1-c} & q_{m2}\,\delta_2^{1-c} \end{pmatrix}$$

$$= \begin{pmatrix} h_{11} & h_{12} \\ \vdots & \vdots \\ h_{j1} & h_{j2} \\ \vdots & \vdots \\ h_{m1} & h_{m2} \end{pmatrix} = \begin{pmatrix} h_1^T \\ \vdots \\ h_j^T \\ \vdots \\ h_m^T \end{pmatrix}$$

기존의 SAS/MARKET 모듈에서는 $c = 0$을 선택하여 행표시자와 열표시자를 각각 $G_{(2)} = (n-1)^{1/2}P_{(2)}$, $H_{(2)} = (n-1)^{-1/2}Q_{(2)}\Delta_{(2)}$로 표현하고 있다(Kuhfeld, 1993).

예제 2.1: 앞에서 설명한 인터넷 검색엔진에 대한 사례연구 자료에 기존의 SAS/MARKET 모듈에서 사용하는 $c = 0$ 을 선택하여 행표시자와 열표시자를 적용시킨 결과그림은 그림 2.1과 같다.

하지만 0에서 1사이의 임의의 c를 선택한다 할지라도 고정된 j번째 그룹에 대해서는 각 상품 i $(i = 1, \cdots, n)$에 대한 선호도 즉, 정사영 거리의 상대적인 비율에는 변함이 없다. 또한 c 값에 상관 없이 각 그룹별로 관측된 선호도와 유도된 2차원 공간에서의 선호도의 상관계수는 다음의 과정을 통하여 동일하다는 것을 알 수 있다. 우선 i상품을 j그룹의 벡터에 직교투사한 정사영의 거리 b_{ij}는 다음과 같으며,

$$b_{ij} = \frac{p_{i1}\,q_{j1}\,\delta_1 + p_{i2}\,q_{j2}\,\delta_2}{\| \,(\, q_{j1}\,\delta_1^{1-c},\; q_{j2}\,\delta_2^{1-c}\,)\,\|} \tag{2.4}$$

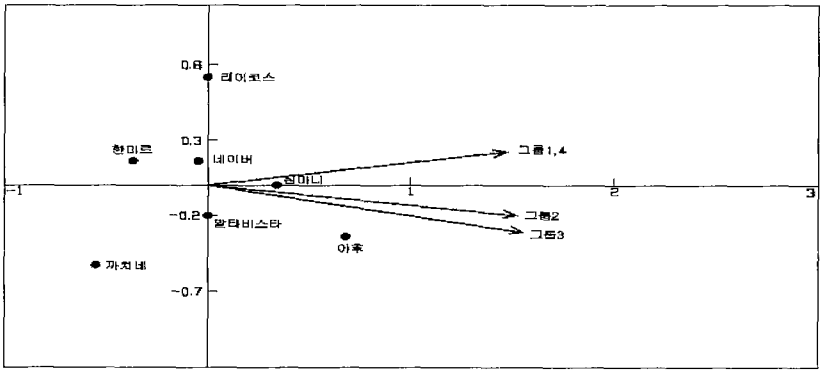

그림 2.1: c=0일 때의 다차원선호도분석

유도된 2차원상에서 j번째 그룹의 n개 상품에 대한 선호도벡터는 b_j로 추정된다.

$$b_j = \begin{pmatrix} \dfrac{p_{11}\, q_{j1}\, \delta_1 + p_{12}\, q_{j2}\, \delta_2}{\| (\, q_{j1}\, \delta_1^{1-c},\ q_{j2}\, \delta_2^{1-c}\,) \|} \\ \vdots \\ \dfrac{p_{n1}\, q_{j1}\, \delta_1 + p_{n2}\, q_{j2}\, \delta_2}{\| (\, q_{j1}\, \delta_1^{1-c},\ q_{j2}\, \delta_2^{1-c}\,) \|} \end{pmatrix} \tag{2.5}$$

그런데 상관계수는 임의의 상수를 더하거나 곱하여도 변하지 않기 때문에 c값이 달라지더라도 상관계수는 일정하게 유지된다. 따라서 축소된 2차원 공간에서의 행좌표와 열좌표를 각각 다음과 같이 나타내도 해석상 문제점은 발생하지 않는다.

$$\begin{aligned} i \text{ 번째 행(상품)의 좌표} &= (\, p_{i1}\, \delta_1^c,\ p_{i2}\, \delta_2^c\,) \\ j \text{ 번째 열(그룹)의 좌표} &= (\, q_{j1}\, \delta_1^{1-c},\ q_{j2}\, \delta_2^{1-c}\,) \end{aligned}$$

Gabriel(1971)은 $c = 0,\ 1/2,\ 1$의 세가지 행렬도의 사용을 제안하였다. 보통 $c = 0$ 인 경우 "주성분 행렬도", $c = 1/2$ 인 경우 "대칭행렬도(symmetric biplot)" 그리고 $c = 1$인 경우는 "JK' 행렬도"라고 부른다. 주성분행렬도는 열간의 관계를 주로 다루는데 비해 JK'행렬도는 행간의 관계에 초점이 맞추어져 있으며 대칭행렬도는 행과 열간의 관계 규명에 유용하다고 볼 수 있다. 이들 행렬도에 관한 대수적, 기하적 성질에 대해서는 최용석(1999)의 2장에 잘 정리되어 있다.

본 연구에서는 c를 선택하는데 있어서 두 가지 최적기준을 제시하고 연관된 "최적"의 행표시자와 열표시자를 유도하고자 한다. 결론을 미리 제시하면, 본 연구의 사례분석 자료의 경우 Gabriel(1971)이 권유한 대칭행렬도와 JK'행렬도가 두 가지 최적 기준에 각각 부합하는 것으로 나타나고 있다. 저자들의 역량이 부족하여 이것이 대수학적으로 정확히 최적기준에 따른 결과물인지는 밝히지 못하였으나, 대칭행렬도와 JK'행렬도의 최적성을 새로운 기준하에서 재조명하였다는 데 의의를 찾을 수 있을 것이다.

이제 최적의 상수 c를 알아내기 위해 두 가지 기준인 총적합도(Total Goodness of Fit)와 총부적합도(Total Stress)를 정의하도록 하자.

3. 두 가지 최적 기준

최적화된 c를 구하는데 있어서 행과 열을 동시에 고려하도록 하며, 2차원 축소로 얻은 적합도(goodness of fit)를 다음의 3가지 측면에서 살펴보자.

첫째, 행과 열을 동시에 고려해보자. 그러면 자료행렬 Z를 $Z_{(2)}$로 축소함에 따라 생기는 적합결여도(measure of lack of fit)는 다음과 같이 측정될 수 있다.

$$\| Z - Z_{(2)} \|^2 = tr\{ (Z - Z_{(2)})^T (Z - Z_{(2)}) \} = \sum_{j=3}^{r} \delta_j^2 \tag{3.1}$$

또한

$$\| Z \|^2 = tr\{ Z^T Z \} = tr\{ (\sum_{j=1}^{r} \delta_j p_j q_j^T)^T (\sum_{j=1}^{r} \delta_j p_j q_j^T) \} = \sum_{j=1}^{r} \delta_j^2$$

이므로 자료행렬 Z 의 적합도는 다음과 같이 정의할 수 있다.

$$GF_{(2)RC} = 1 - \frac{\| Z - Z_{(2)} \|^2}{\| Z \|^2} = \frac{\sum_{j=1}^{2} \delta_j^2}{\sum_{j=1}^{r} \delta_j^2} \tag{3.2}$$

둘째, 행 즉 상품간의 거리를 생각해 보자. 그러면 행 수량화에 의한 적합결여도는

$$\| GG^T - G_{(2)} G_{(2)}^T \|^2 = \| P \Delta^{2c} P^T - P_{(2)} \Delta_{(2)}^{2c} P_{(2)}^T \|^2 = \sum_{j=3}^{r} \delta_j^{4c} \tag{3.3}$$

이고,

$$\| GG^T \|^2 = \sum_{j=1}^{r} \delta_j^{4c} \tag{3.4}$$

이므로 행 수량화에 의한 적합도는 다음과 같다.

$$GF_{(2)R} = 1 - \frac{\| GG^T - G_{(2)} G_{(2)}^T \|^2}{\| GG^T \|^2} = \frac{\sum_{j=1}^{2} \delta_j^{4c}}{\sum_{j=1}^{r} \delta_j^{4c}} \tag{3.5}$$

마지막으로 열 즉 그룹간의 거리를 생각해 보자. 위와 마찬가지 방법으로 열 수량화에 의한 적합도는 다음과 같다.

$$GF_{(2)C} = 1 - \frac{\| HH^T - H_{(2)} H_{(2)}^T \|^2}{\| HH^T \|^2} = \frac{\sum_{j=1}^{2} \delta_j^{4(1-c)}}{\sum_{j=1}^{r} \delta_j^{4(1-c)}} \tag{3.6}$$

(3.5)와 (3.6)은 c값에 의존하나, (3.2)는 c에 무관함을 알 수 있다. 그러므로 (3.2)를 제외한 (3.5)와 (3.6)의 합을 최대화하는 c를 찾는 것은 의미가 있다. 이들의 합을 다음과 같이 총적합도(Total Goodness of Fit: TGF)로 정의하자.

$$TGF_{(2)} = \frac{\sum_{j=1}^{2} \delta_j^{4c}}{\sum_{j=1}^{r} \delta_j^{4c}} + \frac{\sum_{j=1}^{2} \delta_j^{4(1-c)}}{\sum_{j=1}^{r} \delta_j^{4(1-c)}} \tag{3.7}$$

따라서 총적합도 관점에서 최적의 c는 다음 조건을 만족하는 값이 된다.

$$\max_{0 \le c \le 1} TGF_{(2)} \tag{3.8}$$

두 번째 기준으로서, c가 변함에 따라 생기는 여러 행 좌표와 열 좌표들 중에서 원자료의 행간 거리와 열간 거리를 동시에 가장 잘 유지하도록 하는 c를 찾아보자.

표준화된 자료행렬 Z에서 u번째 상품(행)과 v번째 상품(행)간의 유클리드거리를 $d_{u,v}$로 표시하자. 그리고 $u > v$ ($u, v = 1, \cdots, n$)인 모든 가능한 $n(n-1)/2$ 개의 상품(행)간 거리 벡터를 d라고 하자.

$$d = \begin{pmatrix} d_{1,2} \\ \vdots \\ d_{n-1,n} \end{pmatrix} \tag{3.9}$$

또한 축소된 2차원상의 행렬도에서 행표시자 $G_{(2)}$로부터 계산된 u번째 상품(행)과 v번째 상품(행)간의 유클리드 거리를 $d^*_{u,v} = \| g_u^T - g_v^T \|$ 로 표시하고, 마찬가지로 모든 가능한 상품간 거리 벡터 d^*를 다음과 같이 정의하자.

$$d^* = \begin{pmatrix} d^*_{1,2} \\ \vdots \\ d^*_{n-1,n} \end{pmatrix} \tag{3.10}$$

그러면 원자료에서 행간거리인 $d_{u,v}$와 축소된 2차원상에서 행간거리인 $d^*_{u,v}$ 사이의 관계는 다음과 같은 회귀모형으로 나타낼 수 있다.

$$d = \beta_0 + \beta_1 d^* + \epsilon \tag{3.11}$$

여기서

$$y = d,\ \beta = [\, \beta_0,\ \beta_1 \,],\ X = [\, 1 | d^* \,],\ 1 = (\, 1, \cdots, 1 \,)^T$$

로 놓으면 위의 회귀모형은 다음과 같이 표현된다.

$$y = X\beta + \epsilon \tag{3.12}$$

다음의 기호를 이용하면

$$\begin{aligned} SSR &= y^T P_X y\,, \quad P_X = X(\,X^T X\,)^{-1} X^T \\ SSE &= y^T(\,I - P_X\,)\,y \\ SSTO &= y^T y \end{aligned}$$

회귀모형의 설명력에 해당하는 결정계수 R^2 는 다음과 같이 계산된다.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{3.13}$$

그러므로 행 관점에서의 부적합도(Stress)인 ST_R을 다음과 같이 정의하자.

$$\begin{aligned} ST\_R &= 1 - R^2 \\ &= \frac{SSE}{SSTO} = \frac{y^T(I - P_X)y}{y^T y} \end{aligned} \tag{3.14}$$

이제, 그룹(열)간의 거리를 생각해 보자. 행간 거리와 마찬가지로 표준화된 자료행렬 Z에서 s번째 그룹(열)과 t번째 그룹(열)간의 유클리드 거리를 $f_{s,t}$로 표시하자. 그리고 $s > t$ ($s,t = 1,\cdots,m$)인 모든 가능한 $m(m-1)/2$개의 그룹(열)간 거리 벡터를 다음과 같이 f로 놓자.

$$f = \begin{pmatrix} f_{1,2} \\ \vdots \\ f_{m-1,m} \end{pmatrix} \tag{3.15}$$

또한 축소된 2차원상의 행렬도에서 열표시자 $H_{(2)}$로부터 계산된 s번째 그룹(열)과 t번째 그룹(열)간의 유클리드거리를 $f^*_{s,t} = \| h^T_s - h^T_t \|$ 로 표시하고 다음의 기호를 사용하면,

$$\tilde{y} = f,\ \alpha = [\,\alpha_0,\ \alpha_1\,],\ \tilde{X} = [\,1|\,f^*\,],\ 1 = (\,1,\ \cdots,\ 1\,)^T$$

행 관점에서의 부적합도(Stress)인 ST_R과 마찬가지로 열 관점에서의 부적합도 ST_C를 다음과 같이 정의할 수 있다.

$$ST\_C = \frac{\tilde{y}^T\,(\,I - P_{\tilde{X}}\,)\,\tilde{y}}{\tilde{y}^T\,\tilde{y}} \tag{3.16}$$

결국, 행과 열을 동시에 고려한 개념인 총부적합도(Total Stress)인 ST_T를 다음과 같이 정의하는 것이 타당할 것이다.

$$ST\_T = ST\_R + ST\_C = \frac{y^T(I - P_X)y}{y^T y} + \frac{\tilde{y}^T(I - P_{\tilde{X}})\tilde{y}}{\tilde{y}^T\tilde{y}} \tag{3.17}$$

따라서 $\min_{0 \le c \le 1} ST\_T$ 을 만족하는 c가 원자료의 상품간 거리와 그룹간 거리를 동시에 가장 잘 나타내는 2차원상의 행 좌표와 열 좌표를 제공한다.

예제 3.1: 앞의 예제에서 사용한 인터넷 검색엔진 자료에서 (3.7)의 총적합도와 (3.17)의 총부적합도를 기준으로 한 최적의 c를 구해 보자. 먼저, 총적합도를 기준으로 최적의 c를 찾아보면

$$GF_{(2)R} = \frac{\sum_{j=1}^{2} \delta_j^{4c}}{\sum_{j=1}^{r} \delta_j^{4c}}, \quad GF_{(2)C} = \frac{\sum_{j=1}^{2} \delta_j^{4(1-c)}}{\sum_{j=1}^{r} \delta_j^{4(1-c)}}$$

이므로 이 두 함수는 $c = 1/2$ 에 대하여 서로 대칭이다. 또한 표 3.1을 보면 $GF_{(2)R}$은 c에 대한 증가함수이며 오목함수(concave function)이고, $GF_{(2)C}$는 c에 대한 감소함수이며 마찬가지로 오목함수(concave function)이다. 따라서 $GF_{(2)R}$와 $GF_{(2)C}$의 합으로 정의되는 총적합도 $TGF_{(2)}$는 $c = 1/2$ 에서 최대값을 갖게된다. 즉 $TGF_{(2)}$ 관점에서 보면 대칭행렬도가 최적임을 알 수 있다.

표 3.1: c값의 변화에 따른 TGF

c	$GF_{(2)R}$	$GF_{(2)C}$	$TGF_{(2)}$
0.0	0.500	0.999	1.500
0.1	0.651	0.999	1.650
0.2	0.797	0.999	1.796
0.3	0.900	0.997	1.897
0.4	0.956	0.993	1.949
0.5	0.982	0.982	**1.964**
0.6	0.993	0.956	1.949
0.7	0.997	0.900	1.897
0.8	0.999	0.797	1.796
0.9	0.999	0.651	1.650
1.0	0.999	0.500	1.500

그림 3.1에 제시된 $c = 1/2$ 일 때의 다차원선호분석 그림을 보면, 그룹을 나타내는 4개의 벡터는 모두 두 번째 축을 기준으로 오른쪽을 향하고 있다. 이는 조사에 참여한 대부분의 사람들이 야후나 심마니는 선호하는데 비해, 두 번째 축의 왼쪽에 위치하는 까치네, 한미르, 네이버 등은 그다지 선호하지 않음을 의미한다. 검색엔진에 대한 선호도는 그룹1과 그룹4가 비슷하고, 또 그룹2와 그룹3이 비슷한 경향을 나타내고 있다. 선호하는 순위는 조금 차이를 보인다. 즉 그룹1, 2, 3, 4 모두 야후, 심마니 순으로 선호하는 것으로 나타나고 있으나, 그 다음으로 그룹1과 그룹4는 라이코스를 선호하는데 비해, 그룹2와 그룹3은 알타비스타를 선호한다.

이제, 총부적합도를 기준으로 최적의 c를 찾아보자. 표 3.2에 제시된 c 값의 변화에 따른 부적합도를 보면, ST_R은 c가 증가함에 따라 감소하여 $c = 1$일 때 최소값을 갖는다. 즉, $c = 0$에 해당하는 그림 2.1의 기존 SAS/MARKET 모듈에 의한 다차원선호분석 결과에서의 검색엔진(행)간 거리는 원 선호도 자료의 행간거리의 약 87%를 설명하는데 비해,

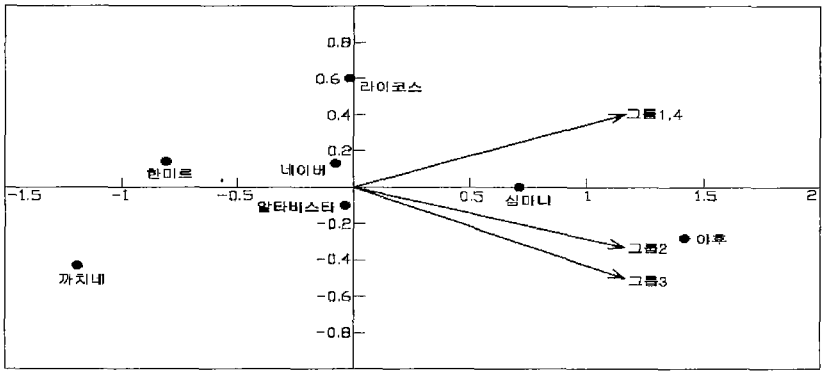

그림 3.1: c=0.5일 때의 다차원선호도분석

$c = 1$일 때의 다차원분석은 약 99%를 설명하고 있다(표 3.2의 ST_R 참조). 이는 두 개의 다차원분석 결과를 비교해보면 알 수 있는데, $c = 0$인 경우에 해당하는 그림 2.1에서는 야후나 심마니가 한미르, 네이버, 알타비스타 등과 비슷한 위치에 존재하여 한 군집을 이루므로 이들 간에 별 차이가 없는 것으로 나타났으나, $c = 1$인 경우에 해당하는 그림 3.2에서는 야후나 심마니가 이들과는 다른 특성을 가져 각각 하나의 독립된 군집을 이루고 있음을 알 수 있다. 표 3.2에서 ST_C는 c가 증가함에 따라 감소하기는 하지만, 그 차이는 거의 없음을 알 수 있다.

표 3.2: c값의 변화에 따른 Stress

c	ST_R	ST_C	ST_T
0.0	0.131	0.0016504	0.133
0.1	0.090	0.0016491	0.092
0.2	0.056	0.0016482	0.057
0.3	0.032	0.0016474	0.033
0.4	0.017	0.0016471	0.018
0.5	0.008	0.0016469	0.010
0.6	0.004	0.0016467	0.006
0.7	0.002	0.0016465	0.003
0.8	0.001	0.0016465	0.002
0.9	0.000	0.0016464	0.002
1.0	0.000	0.0016463	0.002

따라서 ST_R와 ST_C의 합으로 정의되는 총부적합도 ST_T는 $c = 1$일 때 최소가 되며, 이는 ST_T를 기준으로 하면 JK'행렬도가 최적임을 보여 주고 있다. $c = 1$일 때 다른 경우에 비하여 원자료의 검색엔진(행)간 거리와 그룹(열)간 거리를 더 잘 유지하고 있다고 할 수 있다.

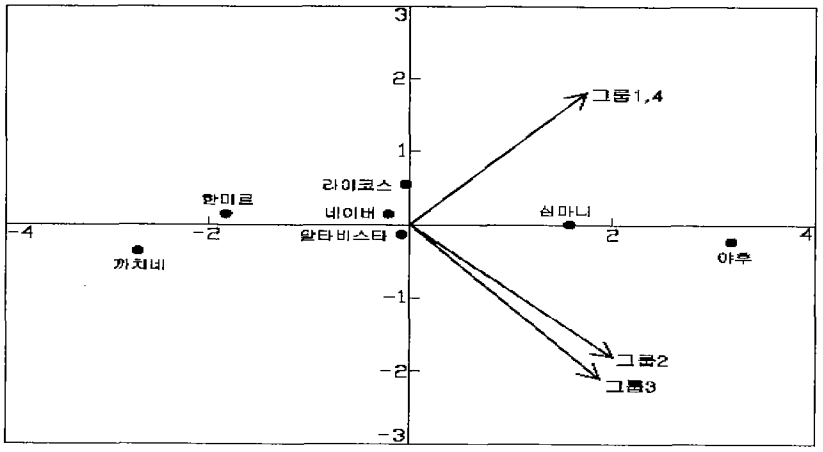

그림 3.2: c=1일 때의 다차원선호도분석

4. 다차원선호분석의 부분수량화

다변량분석기법을 적용시킬 때 분석자의 사전지식이 고려된다면 자료의 분석과 해석에 도움을 얻을 수도 있을 것이다. 자료분석에 연구자의 주관적인 판단을 반영할 수 있다면 자료에 포함된 정보와 사전지식의 접목이 가능하다는 측면에서 심리학, 경영학 등에 특히 유용하리라 판단된다. 이렇듯 사전지식을 다변량분석에 반영하는 과정을 부분수량화(partial quantification)라 부르며 자세한 내용은 서혜선(1999), 허명회(1999) 그리고 황선영과 이연숙(1999)을 참고하기 바란다.

특히 다차원선호분석에서 부분수량화는 연구자가 소비자들의 상품선택에 가장 큰(결정적) 영향을 주는 속성을 알고 있는 경우 사전에 이를 제 1축으로 주고 나머지 축을 찾아내어 분석하는 과정이라 할 수 있다.

분석의 대상이 되는 열별로 표준화한 자료행렬 $Z_{n\times m}$에 대한 부분수량화 과정은 근본적으로 인자분석에서의 부분수량화(서혜선, 1999)와 유사하며 자세한 유도 과정은 부록에 수록하였다.

예제 4.1: 앞에서 사용한 인터넷 검색엔진 자료에 대해 생각해보자. NPD Group Inc.의 1997년 8월 조사에 의하면 대부분(약90%이상)의 사람들이 인터넷 검색엔진 선택시 속도, 정확성, 용이성, 최신정보 보유량의 4가지 속성을 가장 중요하다고 생각하고 있으며 그 중에서도 특히 96%~97%의 사람들은 속도를 '매우 중요하다(Extremely/Very Important)'고 생각하고 있는 것으로 나타났다(source : http://www.npd.com/c_online4.htm). 따라서 인터넷 검색엔진 자료의 경우 속도를 사전 정의된 제1축으로 놓는 부분수량화를 도입한 다차원선호분석방법의 구현이 주요한 관심 대상이 될 수 있다.

이와 같이 인터넷 검색엔진 자료에 부분수량화 과정을 적용시켜, 행렬 Z의 유사비정칙

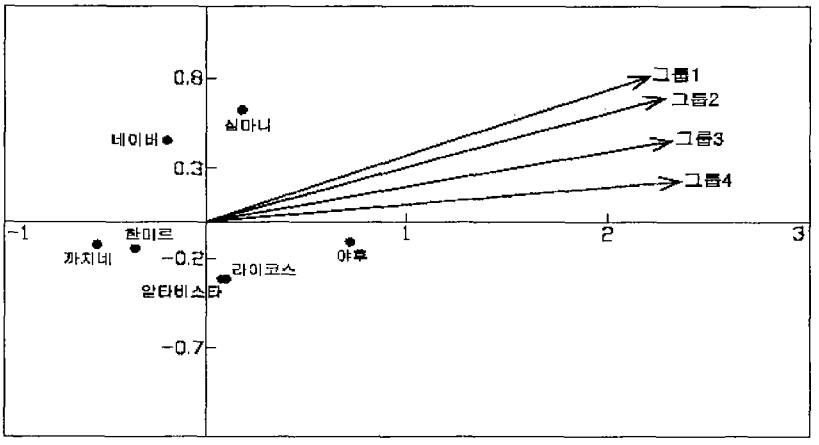

그림 4.1: 부분수량화에 의한 다차원선호도분석

분해에 의해 얻어진 유사고유값 l_j ($j = 1, \cdots, 4$)는 다음과 같다.

$$l_1 = 21.645, \quad l_2 = 1.660, \quad l_3 = 0.492, \quad l_4 = 0.190$$

따라서 2차원으로 축소된 자료행렬의 적합도는 약 97%이며, 이 중 사전에 정의된 제 1축이 약 90%를, 그리고 제 2축이 약 7%를 설명하고 있다.

부분수량화에 의한 결과인 그림 4.1을 보면 제 1축은 속도를 의미하므로 원점을 기준으로 우측에 위치하는 검색엔진들은 속도가 빠른 검색엔진들이고, 좌측에 위치하는 검색엔진들은 속도가 느림을 의미한다(즉 야후, 심마니, 라이코스, 알타비스타, 네이버, 한미르, 까치네 순으로 속도가 빠르다). 또한, 속도를 의미하는 제 1축과 각 그룹을 나타내는 벡터들이 이루는 각을 보아 그룹4, 3, 2, 1 순으로 검색엔진 선택시 속도를 중요하게 고려하고 있음을 알 수 있다.

감사의 글

본 논문의 초고를 총체적으로 수정 보완할 수 있도록 귀중한 조언을 주신 두 분의 심사위원님들께 깊은 감사를 드립니다.

참고문헌

[1] 서혜선(1999). 인자분석에 의한 부분수량화, < 응용통계연구 >, 제12권 1호, 165-173.

[2] 임종원(1997). <마케팅조사 이렇게 한다>, 법문사, 서울.

[3] 최용석(1995). <SAS 다차원척도법>, 자유아카데미, 서울.

[4] 최용석(1999). <행렬도의 이해와 응용>, 부산대학교출판부, 부산.

[5] 황선영, 이연숙(1999). 정준판별분석의 조건부계량화, <응용통계연구> 제12권 2호, 517-525.

[6] 허명회(1999). <다변량 수량화>, 자유아카데미, 서울.

[7] K. R. Gabriel(1971). The Biplot graphic display of matrices with application to principal component analysis, *Biometrika.,* **58**, 453-467.

[8] W.F. Kuhfeld(1993). *Graphical methods for marketing research,* SAS Institute Inc.

[2001년 2월 접수, 2001년 4월 채택]

부록 A: 다차원선호분석의 부분수량화 과정

- 단계 1: 사전 정의된 제 1축을 a_1이라 놓자. 단, a_1은 길이가 1인 단위벡터, 즉 $a_1^T a_1 = 1$이다. 후속수량화를 위해 Z를 다음과 같이 대치한다.

$$Z_{(1)} = (I_n - P_{a_1}) Z \tag{A.1}$$

여기서 $P_{a_1} = a_1 (a_1^T a_1)^{-1} a_1^T$ 이다.

- 단계 2: 임의의 $m \times 1$ 단위벡터 $c_2 = (c_1, \cdots, c_m)^T$에 대해

$$max \| Z_{(1)} c_2 \|^2$$

을 만족하는 c_2를 구한 후, 이를 이용하여 다음 제 2축을 얻는다.

$$a_2 = \frac{Z_{(1)} c_2}{\| Z_{(1)} c_2 \|}$$

이제 (단계 1)에서 도출된 $Z_{(1)}$을 다음과 같이 대치한다

$$Z_{(2)} = (I_n - P_{a_2}) Z_{(1)} \tag{A.2}$$

- 단계 3: 이러한 과정을 m번 수행하여 나머지축 $a_3, \cdots, a_m$을 얻는다.

한편 위의 부분수량화 과정을 만족하는 c_j ($j = 1, \cdots, m$)는 다음의 고유체계를 통해 구할 수 있다(허명회, 1999).

$$Z_{(1)}^T Z_{(1)} c_j = \lambda_j c_j , \lambda_2 \geq, \cdots, \geq \lambda_m$$

부분수량화 과정의 (단계 1)에서 (A.1)에 의하면 $Z = a_1 a_1^T Z + Z_{(1)}$으로 표현되며, (단계 2)의 (A.2)로부터

$$Z_{(2)} = (I_n - P_{a_2}) Z_{(1)} = Z - a_1 a_1^T Z - a_2 a_2^T Z_{(1)}$$

이 때, $Z_{(1)}^T a_2 = Z^T a_2$ 이므로 $Z_{(2)}$는 다음과 같이 표현될 수 있다.

$$Z_{(2)} = Z - a_1 a_1^T Z - a_2 a_2^T Z$$

따라서, $Z_{(3)}, \cdots, Z_{(m)}$등을 계속 적용하면 다음 결과를 얻는다.

$$Z = a_1 a_1^T Z + a_2 a_2^T Z + \cdots + a_m a_m^T Z + Z_{(m)}$$

그런데 $Z_{(m)} = 0$ 이므로 열별로 표준화한 자료행렬 $Z_{n \times m}$은 다음과 같다.

$$Z = \sum_{j=1}^{m} a_j a_j^T Z$$

한편, l_j와 c_j를 각각 다음과 같이 정의하자.

$$l_j = a_j^T \, ZZ^T \, a_j \, , \quad j = 1, \cdots, m$$

$$c_j = \frac{Z^T a_j}{\sqrt{l_j}}, \quad j = 1, \cdots, m$$

그러면,

$$Z = \sum_{j=1}^{m} \sqrt{l_j} \, a_j \, c_j^T = A \, D_{\sqrt{l}} \, C^T \tag{A.3}$$

여기서 $A = (a_1, \cdots, a_m)$, a_j는 j축 벡터

$$D_{\sqrt{l}} = diag (\sqrt{l_1}, \cdots, \sqrt{l_m}), \quad C = (c_1, \cdots, c_m)$$

이다. (A.3)의 A 와 C 는 다음을 만족한다.

$$A^T A = I_m \, , \quad C^T C = \begin{pmatrix} 1 & n^T \\ n & I_{m-1} \end{pmatrix}$$

여기서 $n = (n_2, \cdots, n_m)$, $n_j = c_1^T \, c_j$ ($j = 2, \cdots, m$)이다. (A.3)을 행렬 Z의 유사비정칙값분해라 하고, l_j를 유사고유값이라 한다(서혜선, 1999). 유사비청칙값분해에 의하여 Z는 다음과 같이 표현할 수 있다.

$$Z = A \, D_{\sqrt{l}} \, C^T = G \, H^T$$

여기서 행 좌표와 열 좌표는 다음과 같이 구할 수 있다.

$$G = A \, (\text{행 표시자}), \quad H = C \, D_{\sqrt{l}} \, (\text{열 표시자})$$

이를 s차원으로 축소하여 근사적으로 표현하면 다음과 같다.

$$Z_{(s)} = A_{(s)} \, D_{\sqrt{l}(s)} \, C_{(s)}^T = G_{(s)} \, H_{(s)}^T$$

여기서, $G_{(s)}$와 $H_{(s)}$는 각각 행렬 G와 H의 처음 s개의 열을 갖는다. 이와 같이 s차원으로 축소된 자료행렬의 적합도(goodness of fit)는 다음과 같다(최용석, 1999).

$$\frac{\| G_{(s)} H_{(s)}^T \|^2}{\| GH^T \|^2} = \frac{tr(H_{(s)} G_{(s)}^T G_{(s)} H_{(s)}^T)}{tr(HG^T GH^T)} = \sum_{j=1}^{s} l_j \; / \; \sum_{j=1}^{m} l_j \tag{A.4}$$

(A.4)는 다음 관계식에 의해 유도 된 것이다.

$$\begin{aligned} tr(HG^TGH^T) &= tr(HH^T) \ (\because G^TG = A^TA = I_m) \\ &= tr(\ D_{\sqrt{l}}\ C^TCD_{\sqrt{l}}\ ) \\ &= tr(\ D_{\sqrt{l}}\ \begin{pmatrix} 1 & n^T \\ n & I_{p-1} \end{pmatrix} D_{\sqrt{l}}\ ) \\ &= \sum_{j=1}^{m} l_j \end{aligned}$$

Optimal Scaling and Partial Quantification in Multidimensional Preference Analysis

S.Y. Hwang [1] Su-Jin Jung [2] Young-Won Kim [3]

ABSTRACT

Multidimensonal preference analysis(MDPREF) is an useful analytical tool which figures out graphically how consumers recognize, evaluate and prefer certain products. This article mainly concerns the optimal scaling and partial quantification in implementing MDPREF. Methodologies developed here can be directly applicable to the real data. As an example, internet searching engine data in Korea are analized for implementing and illustrating the main results of the article.

Keywords: Biplot; Multidimensonal preference analysis; Optimal scaling; Partial quantification.

1) Associate Professor, Dept. of Statistics, Sookmyung Women's Univ.
E-mail: shwang@sookmyung.ac.kr
2) Graduate Student, Dept. of of Statistics, Sookmyung Women's Univ.
3) Professor, Dept. of Statistics, Sookmyung Women's Univ.
E-mail: ywkim@sookmyung.ac.kr