The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 651-659

Preference Map using Weighted Regression[1)]

S. Y. Hwang[2)], Su-Jin Jung[3)], Young-Won Kim[4)]

Abstract

Preference map is a widely used graphical method for the preference data set which is frequently encountered in the field of marketing research. This provides joint configuration usually in two dimensional space between "products" and their "attributes". Whereas the classical preference map adopts the ordinary least squares method in deriving map, the present article suggests the weighted least squares approach providing the better graphical display and interpretation compared to the classical one. Internet search engine data in Korea are analysed for illustration.

Keywords : Attribute, Preference map, Weighted least squares method

1. 서론

선호도(preference map)는 "상품"들과 이들 상품의 "속성(attribute)"을 선호값 또는 순위값를 이용하여 저차원(보통 2차원이므로 본 논문에서는 2차원만을 고려하기로 한다) 공간에 상품과 속성을 동시에 그림으로 표시하는 분석기법으로서 공통 공간에 동시에 나타낼 때 상품은 점으로, 속성은 벡터로 나타낸다(Kuhfeld, 1993). 선호도 분석에 이용되는 선호값(순위값)은 큰 값일수록 더욱 선호되고 더 좋은 순위를 표시한다. 선호도는 원래 Biplot(Gabriel, 1971; 최용석, 1999)의 응용기법이며 마케팅리서치분야에서 널리 이용되는 방법으로서(임종원, 1997) 자료의 SVD(Singular Value Decimposition)를 이용한 저차원 근사화가 핵심 내용이며, SAS에서는 MARKET 메뉴를 이용해 선호도를 작성할 수 있다.

기존의 선호도에서의 속성 벡터는 보통최소제곱법(ordinary least squares method)을 이용하여 구하고 있으나 본 연구에서는 가중최소제곱법(weighted least squares method)을 이용한 선호도 작성을 제안하고 있다. 특히 제안된 방법은 소비자 그룹별 선호도들의 비교분석을 더욱 정확하게 해준다. 2절에서는 기존 선호도 이론을 요약하고 있으며, 제안된 가중최소제곱법을 3절에서 소개하고, 인터넷 검색엔진 자료에 적용시킨 사례분석 결과는 4절에 수록하였다.

---

1) This work was supported by a grant from Korea Research Foundation(KRF-2000-015-DP0052).

2) Associate Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail : shwang@sookmyung.ac.kr

3) Graduate, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.

4) Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail : ywkim@sookmyung.ac.kr

2. 선 호 도

먼저 기존의 선호도를 작성하는데 필요한 수식을 살펴보도록 하자. n개의 상품이 이차원상에 좌표로 주어졌을 때(이 좌표들은 보통 다차원척도법 또는 다른 다변량분석기법(예를 들어 요인분석, 다차원 선호분석)을 통해 유도될 수 있다(Kuhfeld, 1993).), 이 좌표들을 $n\times 2$ 행렬 $P_{(2)}$로 표시하고 상품속성을 나타내는 크기가 $n\times 1$인 w개의 속성변수 $\boldsymbol{y}_k$ $(k=1,\cdots,w)$를 생각해 보자. 여기서 $\boldsymbol{y}_k$는 n개의 상품에 대한 선호값으로 구성된 벡터이다. 또한 $\boldsymbol{B}_{2\times w}$이 2차원상에서 w개의 속성을 나타내는 벡터로 구성된 행렬일 때, 다음의 다변량 선형회귀모형을 생각하자.

$$Y_{n\times w} = P_{(2)}\ B_{2\times w} + \varepsilon_{n\times w} \tag{2.1}$$

여기서, $P_{(2)}$는 2차원상의 상품좌표로 구성된 행렬이며 $\boldsymbol{\beta}_k$는 k번째 속성벡터이고, 즉,

$$Y_{n\times w} = [\ \boldsymbol{y}_1 |\ \boldsymbol{y}_2 | \cdots |\ \boldsymbol{y}_w\ ]$$

$$B_{2\times w} = [\ \boldsymbol{\beta}_1 |\ \boldsymbol{\beta}_2 | \cdots |\ \boldsymbol{\beta}_w\ ]$$

$$\varepsilon_{n\times w} = [\ \boldsymbol{\varepsilon}_1 |\ \boldsymbol{\varepsilon}_2 | \cdots |\ \boldsymbol{\varepsilon}_w\ ]$$

을 나타내며 $E(\ \boldsymbol{\varepsilon}_k) = \mathbf{0}$ 이고, $Cov(\ \boldsymbol{\varepsilon}_k,\ \boldsymbol{\varepsilon}_h) = \delta_{kh} I$, 여기서 $k, h = 1,\cdots,w$ 임을 가정한다. 즉, i번째 상품 $(i=1,\cdots,n)$에 대해 w개의 속성값들의 분산공분산은 $\Delta = \{\delta_{kh}\}$로 상품의 종류에 관계없이 동일하며, 서로 다른 상품에 대한 속성값들은 서로 독립적이다.

따라서 (2.1) 모형에서 보통최소제곱법에 의해 추정된 모수는 다음과 같다,

$$\widehat{B}_{2\times w} = [\ \widehat{\boldsymbol{\beta}}_1 |\ \widehat{\boldsymbol{\beta}}_2 | \cdots |\ \widehat{\boldsymbol{\beta}}_w\ ] \quad = (P_{(2)}^T P_{(2)})^{-1} P_{(2)}^T Y \tag{2.2}$$

결국 k번째 속성을 나타내는 벡터는 단변량 선형회귀모형,

$$\boldsymbol{y}_k = P_{(2)}\ \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k \tag{2.3}$$

으로부터 보통최소제곱법으로 다음과 같이 추정된다.

$$\widehat{\boldsymbol{\beta}}_k = (P_{(2)}^T P_{(2)})^{-1} P_{(2)}^T\ \boldsymbol{y}_k \tag{2.4}$$

이제 본 연구에서 다루고자 하는 상황인, 전체 표본이 몇 개의 그룹으로 구성된 경우를 생각해 보자(이런 경우는 실제 자료분석에서 자주 나타나며 4절에서 다루고 있는 사례연구 자료가 이에 해당한다). 이런 경우 보통은 모든 표본 구성원의 선호도를 평균하여 전체표본에 대한 하나의 선호도를 작성하고 있다. 이러한 분석은 각 그룹에서 상품들에 대한 선호하는 정도가 동질적이라는 가정에 근거를 두고 있다. 본 논문에서는 그룹별 선호정도의 동질성에 의문을 던지고 해결방안으로서 가중최소제곱법을 제안하고 있다. 또한 각 그룹별로 선호도를 구해 서로 비교해 보기 위해서 그룹별 선호도를 이분산성 개념을 도입해 작성해 보고자 한다.

3. 가중회귀를 이용한 선호도

그룹별 선호도의 이질성을 반영하기 위해 j 그룹 $(j=1,\cdots,m)$에서 임의의 속성 a 에 대한 선

호도를 생각해보자. j 그룹은 u_j명으로 구성되어 있고, $y_{jil}\,(i=1,\cdots,n;\ l=1,\cdots,u_j)$을 i번째 상품의 속성 a에 대한 j 그룹의 u_j명의 선호도값이라 하면, 이에 해당하는 자료 행렬은 다음과 같다.

$$\{y_{jil}\} = \begin{pmatrix} y_{j11} & \cdots & y_{j1l} & \cdots & y_{j1u_j} \\ \vdots & & \vdots & & \vdots \\ y_{ji1} & \cdots & y_{jil} & \cdots & y_{jiu_j} \\ \vdots & & \vdots & & \vdots \\ y_{jn1} & \cdots & y_{jnl} & \cdots & y_{jnu_j} \end{pmatrix} \tag{3.1}$$

한편, 속성 a에 대해 j그룹에서 n개 상품에 대한 평균 선호도값 벡터를 다음과 같이 $\boldsymbol{y}(j)$로 표현하고,

$$\boldsymbol{y}(j) = \begin{pmatrix} \bar{y}_{j1.} \\ \vdots \\ \bar{y}_{ji.} \\ \vdots \\ \bar{y}_{jn.} \end{pmatrix}, \text{ 여기서 } \bar{y}_{ji.} = \frac{\sum_{l=1}^{u_j} y_{jil}}{u_j} \tag{3.2}$$

$\boldsymbol{y}(j)$를 평균과 분산이 각각 0과 1이 되도록 표준화시킨 벡터를 $\boldsymbol{y}^*(j)$라 하자. 그러면 다음과 같은 회귀모형을 통해 j그룹에 대한 속성 a의 2차원상의 좌표에 해당하는 $\boldsymbol{\beta}(j)$에 대한 추정값을 얻을 수 있다.

$$\boldsymbol{y}^*(j) = P_{(2)}\,\boldsymbol{\beta}(j) + \boldsymbol{\varepsilon}(j) \tag{3.3}$$

기존의 선호도에서는 $Var(\boldsymbol{y}^*(j)) = Var(\boldsymbol{\varepsilon}(j)) = \sigma^2 I_{n\times n}$을 가정하였다. 즉, 각 상품마다 선호값의 분산이 동일하다고 가정한 것이다. 그러나 실제 자료에서는 j번째 그룹에서 각 상품 $(i=1,\cdots,n)$에 대한 선호값의 흩어짐 정도 즉, 분산이 다를 수 있다. 다시 말해 i번째 상품에 대한 u_j명의 선호값의 분산과 $i'\ (i\neq i')$번째 상품에 대한 u_j명의 선호값의 분산은 다를 수가 있다. 즉, $\sigma_{ji}^2 \neq \sigma_{ji'}^2\ (i\neq i')$을 가정하는 것이 보다 현실적이라 할 수 있다. 따라서 j그룹의 i상품에 대한 분산인 σ_{ji}^2을 다음과 같은 형태로 표현해 보자.

$$\sigma_{ji}^2 = 1/w_{ji}, \quad W_j = \begin{pmatrix} w_{j1} & & 0 \\ & \ddots & \\ 0 & & w_{jn} \end{pmatrix} \tag{3.4}$$

가중회귀방법은

$$Q_{jw} = (\boldsymbol{y}^*(j) - P_{(2)}\,\boldsymbol{\beta}(j))^T W_j\,(\boldsymbol{y}^*(j) - P_{(2)}\,\boldsymbol{\beta}(j))$$

을 최소화하는 과정이며, 가중 최소제곱정규방정식은 다음과 같다.

$$(P_{(2)}^T W_j P_{(2)})\,\boldsymbol{\beta}(j) = P_{(2)}^T W_j\,\boldsymbol{y}^*(j) \tag{3.5}$$

따라서 j그룹의 속성 a 벡터의 2차원상의 좌표는

$$\hat{\boldsymbol{\beta}}(j) = (P_{(2)}^T W_j P_{(2)})^{-1} P_{(2)}^T W_j \boldsymbol{y}^*(j) \tag{3.6}$$

으로 추정된다.

현실적으로 σ_{ji}^2을 알 수 없기 때문에 이를 추정할 필요가 있으며, w_{ji}은 i번째 상품에 대한 j그룹 u_j명의 속성 a에 대한 선호값의 표본분산 (s_{ji}^2)을 사용하여 추정할 수 있다. 따라서 표본분산,

$$s_{ji}^2 = \frac{\sum_{l=1}^{u_j} (y_{jil} - \bar{y}_{ji})^2}{u_j - 1} \tag{3.7}$$

을 이용하여 $\hat{W}_j$를 다음과 같이 구할 수 있다.

$$\hat{W}_j = \begin{pmatrix} \hat{w}_{j1} & & 0 \\ & \ddots & \\ 0 & & \hat{w}_{jn} \end{pmatrix}, \quad \text{여기서} \quad \hat{w}_{ji} = \frac{1}{s_{ji}^2} \tag{3.8}$$

결과적으로 j그룹의 속성 a 벡터의 2차원상의 좌표를 구하면 다음과 같다.

$$\hat{\boldsymbol{\beta}}(j) = ((P_{(2)}^T \hat{W}_j P_{(2)})^{-1} P_{(2)}^T \hat{W}_j \boldsymbol{y}^*(j) \tag{3.9}$$

이제, 사례분석자료를 통해 제안된 식(3.9)에 의한 결과와 기존의 방법인 식(2.4)에 의한 결과를 비교하고자 한다.

4. 사례연구 및 결론

실증적인 사례분석을 위해 이용한 자료는 검색엔진에 대한 선호자료이다. 매일경제 여론조사(1999년 9월), 마이넷 인터넷 설문조사(1998년 12월 15일~1999년 1월 20일) 그리고 kissnet (1998년 4월 7일)의 설문조사 결과를 참고로 하여 가장 많이 사용되는 7개의 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(http://www.npd.com/c_online4.htm). 설문조사는 인터넷 리서치 전문회사인 kissnet에 의뢰하여, kissnet 홈페이지(http://www.kissnet.co.kr)에서 1999년 9월 17일에서 20일까지 4일에 걸쳐 총 200명에게 이루어졌다.

또한 검색엔진에 대한 선호값은 연령별로 차이가 있을 것으로 예상하여 모집단을 초/중/고등학생, 대학(원)생, 30대 이전의 직장인, 30대이후의 직장인의 4그룹으로 층화하였다.(본 논문에서는 이를 각각 그룹1, 그룹2, 그룹3, 그룹4로 표현한다.) 각 층별로는 50명씩 균등 배분하여 추출하였다. 이렇게 얻은 총 200명의 자료 중 검색엔진별로 선호도의 차이가 전혀 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, 30대이후의 직장인 38명이다. 본 설문조사에서 사용한 설문 내용은 부록에 수록하였다.

이 경우 선호도 결과는 이차원 상에 7개의 상품이 점으로 위치하고 4개의 속성들이 각각 벡터

로 나타나며 "상품"을 "속성"벡터에 직교투사(orthogonal projection)시킨 거리가 추정된 선호값으로 해석된다.

각 그룹별로 식(2.4)의 보통최소제곱법과 식(3.9)의 가중최소제곱법을 이용하여 구한 속성벡터는 <표 2>에 수록하였으며, 가중최소제곱법을 적용시켜 얻은 검색엔진 자료의 각 그룹별 선호도는 <그림 4.1>-<그림 4.4>에 제시하였다. <표 2>를 보면 각 그룹별로 가중최소제곱법 결과는 기존의 보통최소제곱법 결과와 전체적으로 크기나 부호에 있어서 큰 차이가 나지는 않지만 그룹별 선호도에 이분산성 개념을 도입했다는 측면에서 좀더 정확한 추정값을 제공하고 있다고 할 수 있다

<표 2> 각 그룹별 속성벡터(β)의 추정결과

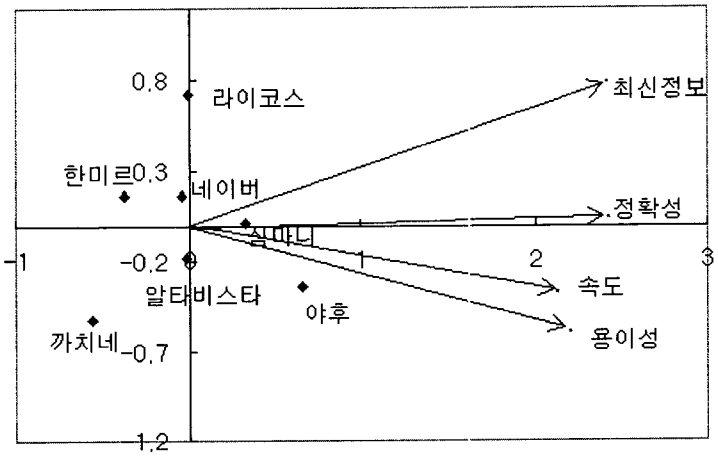
	속성	보통최소제곱법	가중최소제곱법
그룹 1	속도	(2.1939, -0.3907)	(2.1342, -0.3729)
	정확성	(2.4198, 0.0333)	(2.4252, 0.0382)
	용이성	(2.3098, -0.5834)	(2.2095, -0.5970)
	최신정보	(2.2549, 0.8242)	(2.4233, 0.7803)
그룹 2	속도	(2.3224, -0.1418)	(2.3397, -0.1409)
	정확성	(2.3495, -0.1479)	(2.3935, -0.1371)
	용이성	(2.3985, -0.4129)	(2.4022, -0.4129)
	최신정보	(2.1788, 0.5672)	(2.2224, 0.4038)
그룹 3	속도	(2.4086, -0.0267)	(2.4017, -0.0683)
	정확성	(2.4005, -0.3997)	(2.3943, -0.4076)
	용이성	(2.3690, -0.0783)	(2.4593, -0.1109)
	최신정보	(2.4158, -0.1211)	(2.4738, -0.1082)
그룹 4	속도	(2.2646, 0.6818)	(2.2615, 0.7193)
	정확성	(2.4160, 0.1768)	(2.4230, 0.1776)
	용이성	(2.3535, 0.5599)	(2.3615, 0.5747)
	최신정보	(2.2177, 0.7928)	(2.2510, 0.7762)

한편 그룹별 선호도를 살펴보면, 그룹1의 선호도 <그림 4.1>에서 제1축과 정확성을 나타내는 축과의 사이 각이 작으므로 제 1축은 정확성으로 해석할 수 있다. 마찬가지로 그룹2의 선호도 <그림 4.2>에서 제1축은 정확성 또는 속도로, 그룹3의 선호도 <그림 4.3>에서는 속도로, 그리고 그룹4의 선호도 <그림 4.4>에서는 정확성으로 해석할 수 있다.

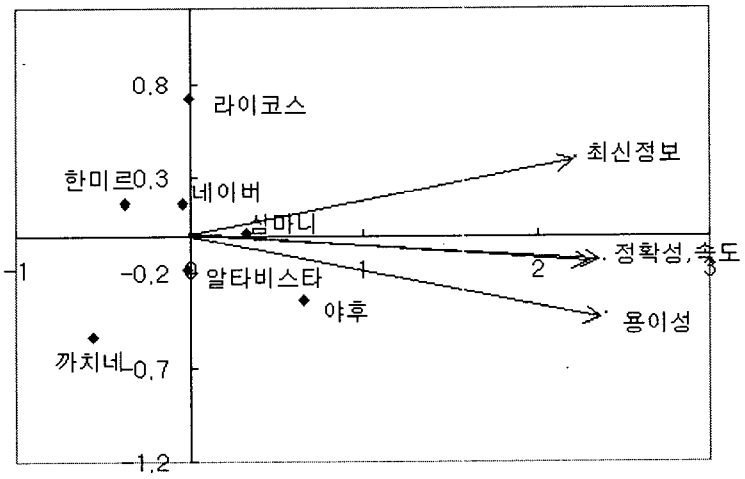
기존의 SAS/MARKET/MDPREF를 설문 자료에 적용시켜 얻은 다차원선호분석 결과인 그룹별 상품 선호도 <그림 4.5>에서 그룹1을 나타내는 벡터와 그룹1의 선호도 <그림 4.1>에서의 최신정보와 정확성을 나타내는 벡터의 사이 각이 다른 속성들에 비해 작음을 알 수 있다. 즉, 그룹1의 사람들이 검색엔진 선택시 중요시 여기는 속성은 최신정보 보유량과 정확성임을 알 수 있다. <그림 4.2>-<그림 4.4>를 참고로 하면 검색엔진 선택시 그룹2는 사용법의 용이성을, 그룹3과 그룹4는 정확성을 다른 속성들보다 중요하게 여기고 있다고 결론지을 수 있다. 또한 그룹2는 속도와 정확성의 상관관계가 높다고 생각하고 있으며, 그룹3은 속도, 용이성, 최신정보들간의 상관관계가 높다는 것을 파악할 수 있다.

이러한 해석은 전체표본을 종합하여 작성되는 기존의 선호도로는 얻을 수 없는 결과이며, 각 그룹별 선호도를 작성하는데 본 연구에서 제시한 가중 최소제곱법을 이용하여 통계적으로 좀더 정

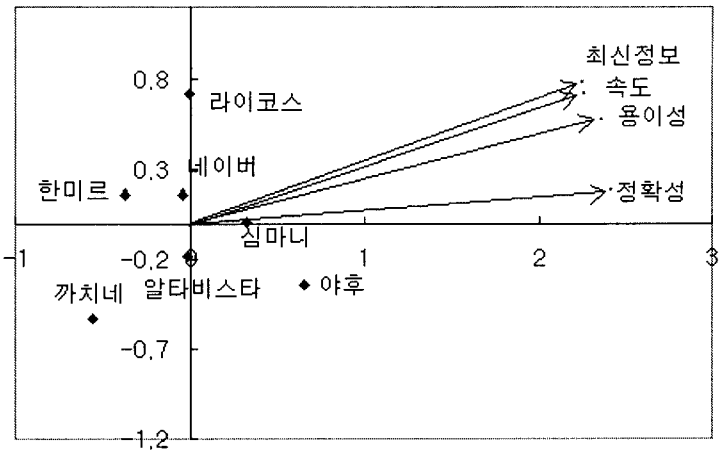
확한 결과를 얻을 수 있다는 점에서 의의를 찾을 수 있겠다. 특히 각 그룹별 이질성이 뚜렷한 자료인 경우에는 기존의 선호도 보다 효율적인 그룹별 선호도를 얻을 수 있어 심층적인 해석이 가능해진다.



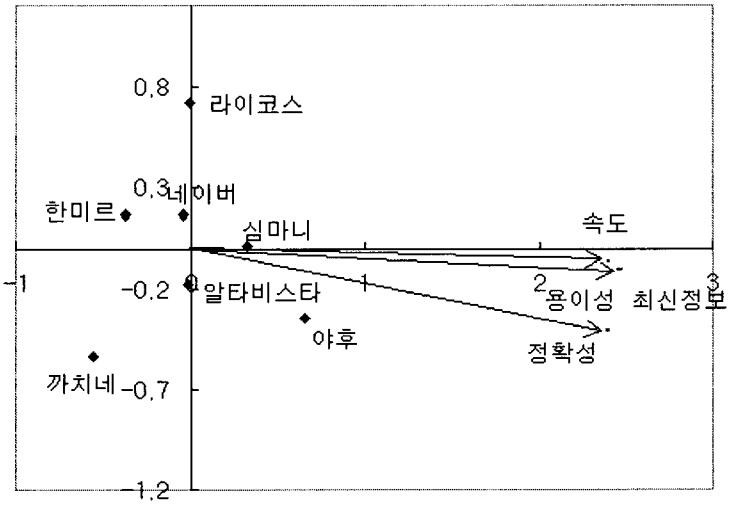
<그림 4.1> 가중최소제곱법을 이용한 그룹1의 선호도



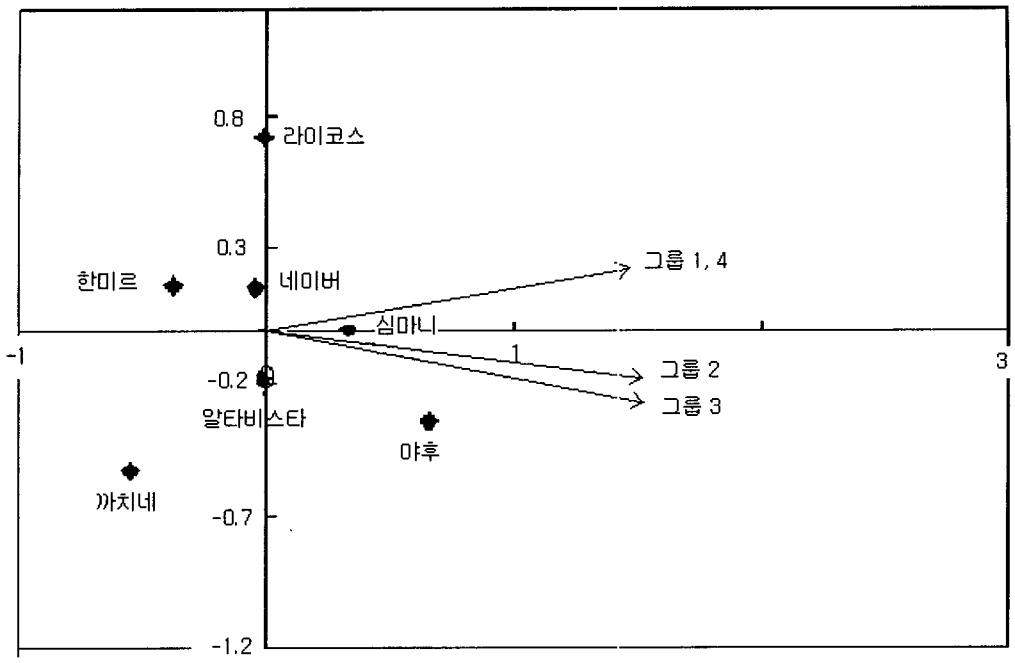
<그림 4.2> 가중최소제곱법을 이용한 그룹2의 선호도



<그림 4.3> 가중최소제곱법을 이용한 그룹3의 선호도



<그림 4.4> 가중최소제곱법을 이용한 그룹4의 선호도



<그림 4.5> 그룹별 상품 선호도

감사의 글

사례연구(4절) 부분을 대폭 수정할 수 있도록 조언해 주신 두 분의 심사위원께 감사 드립니다.

참고문헌

[1] 성내경 (1994). 「SAS/IML-행렬연산」, 자유아카데미, 서울.
[2] 임종원 (1997). 「마케팅조사 이렇게 한다」, 법문사, 서울.
[3] 최용석 (1999). 「행렬도의 이해와 응용」, 부산대학교 출판부, 부산
[4] 허명회 (1999). 「다변량 수량화」, 자유아카데미, 서울.
[5] Gabriel, K.R.(1971). The Biplot graphic display of matrics with application to principal component analysis, *Biometrika,* Vol 58, 453-467.
[6] Johnson, R.A. and Wichern, D.W.(1991). *Applied multivariate statistical analysis.* 3rd. ed.. Prentice Hall.
[7] Kuhfeld, W.F.(1993). *Graphical methods for marketing research.* SAS Institute Inc.

부록: 설문지

[A] 다음은 7가지 검색엔진에 대한 귀하의 선호하는 정도를 알아보기 위한 질문입니다. 각 검색엔진별로 선호하는 정도를 1(전혀 선호하지 않음)에서 9(매우 선호함)까지의 척도를 기준으로 알맞다고 느끼시는 숫자에 표시하여 주시기 바랍니다.

	전혀 선호 안함				보통				매우 선호함
까치네	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
네이버	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
라이코스코리아	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
심마니	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
야후코리아	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
한미르(구정보탐정)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
한글알타비스타	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)

[B] 다음은 7종류의 검색엔진에 대한 귀하의 느낌이나 평가를 각 속성(특성)별로 알아보고자 하는 질문입니다. 각 속성에 대한 평가는 검색엔진별로 하여 주시기 바랍니다.

1. 속도가 빠르다.(검색결과가 빠르게 출력되고 페이지 이동이 빠르다)

	매우 느리다				보통				매우 빠르다
까치네	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
. . .	. . .				. . .				. . .

2. 검색결과가 정확하고 관련성이 높다.

	전혀 정확하지 않음				보통				매우 정확하다
까치네	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
. . .	. . .				. . .				. . .

3. 사용하기 쉽다.

	매우 어렵다				보통				매우 쉽다
까치네	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
. . .	. . .				. . .				. . .

4. 정보가 잘 정리되어 있고 최신정보를 많이 지니고 있다.

	최신 정보가 전혀 없다				보통				정보가 매우 많다
까치네	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)