The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.1.067
31(1), 67–75

Multidimensional scaling of categorical data using the partition method

Sang Min Shin[a] · Sun-Kyung Chun[b] · Yong-Seok Choi[b,1]

[a]Department of Management Information Systems, Dong-A University;
[b]Department of Statistics, Pusan National University

(Received October 23, 2017; Revised December 5, 2017; Accepted December 13, 2017)

Abstract

Multidimensional scaling (MDS) is an exploratory analysis of multivariate data to represent the dissimilarity among objects in the geometric low-dimensional space. However, a general MDS map only shows the information of objects without any information about variables. In this study, we used MDS based on the algorithm of Torgerson (*Theory and Methods of Scaling*, Wiley, 1958) to visualize some clusters of objects in categorical data. For this, we convert given data into a multiple indicator matrix. Additionally, we added the information of levels for each categorical variable on the MDS map by applying the partition method of Shin *et al.* (*Korean Journal of Applied Statistics*, **28**, 1171–1180, 2015). Therefore, we can find information on the similarity among objects as well as find associations among categorical variables using the proposed MDS map.

Keywords: pooling the independent cohort data sets, benchmark dose lower limit, linear mixed model, attention deficit hyperactivity disorder, blood lead level

1. 서론

다차원척도법(multidimensional scaling; MDS)이란 개체간의 비유사성(dissimilarity)을 저차원 공간에 기하적으로 나타내는 다변량 자료에 대한 탐색적 분석기법이다. 여기서, 개체간의 비유사성을 다차원척도법에 의해 투영시키고자 하는 저차원 공간을 형상공간(configuration space)이라 하고 형상공간에 나타낸 그림을 다차원척도그림(MDS map)이라 한다 (Choi, 2014). Cox와 Cox (2000)에 따르면 일반적으로 다차원척도법은 계량형 다차원척도법(metric MDS)과 비계량형 다차원척도법(non-metric MDS)으로 구분할 수 있는데, 이 중 계량형 다차원척도법은 일반적으로 계량형자료에 대해 적용 가능하나, 본 연구에서는 범주형자료에 대해 계량형 다차원척도법을 적용하고자 한다.

더불어 계량형 다차원척도법은 형상공간에 자료행렬의 개체들에 대한 정보만을 표현할 뿐 변수에 대한 정보는 표현하지 못하는 단점이 있다. 이러한 단점을 보완하기 위해 Gower와 Harding (1988)은 비선형행렬도(nonlinear biplot)를 제안하였고 이를 응용하여, Gower (1992)는 계량형변수와 범주형변수를

This work was supported by a 2-Year Research Grant of Pusan National University.
[1]Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea. E-mail: yschoi@pusan.ac.kr

모두 포함한 일반적인 자료에 대해 적용 가능한 일반화행렬도(generalized biplot)를 제안한 바 있다. 이들은 모두 개별 변수의 특성을 파악하기 위한 가상점들(pseudo-points)을 생성하고 이들 가상점들의 중심위치를 파악하여 해당 변수의 정보를 개체간의 비유사성을 표현한 다차원척도법의 형상공간에 투영시키는 방식이다. 이들 방법에 따르면 계량형변수의 정보는 선형 혹은 비선형의 궤적(trajectory)으로 표현되며, 범주형변수의 정보는 범주수준별로 하나의 점(point)으로 표현된다. 그리고 Shin 등(2015)은 이러한 일반화행렬도의 기법을 응용하여 이진수자료(binary data)에 대한 계량형 다차원척도그림 상에 변수의 정보를 표현하는 분할법(partition method)을 제안한 바 있다. 그러나 일반적인 범주형자료는 개별 변수의 범주수준이 2개 이상이므로 본 연구에서는 일반적인 범주형자료에 분할법을 적용하여 범주수준별 가상점을 다차원척도그림에 추가하는 방법을 제안하고, 이를 이용하면 변수와 개체간의 해석이 보다 용이해짐을 보여주고자 한다.

이에 2절에서는 범주형자료에서 계량형 다차원척도법 적용 과정과 분할법에 의한 범주형변수의 정보를 다차원척도법의 형상공간 상에 표현하는 방법을 설명하고 3절에서 활용 사례를 제시한 후, 끝으로 4절에서 정리·요약하려 한다.

2. 분할법을 활용한 범주형자료의 다차원척도법

2.1. 범주형자료의 다차원척도법

n개의 개체들에 대하여 p개의 범주형변수를 측정한 자료행렬 $\mathbf{X} = \{x_{ik}\}$, $i = 1, \ldots, n$, $k = 1, \ldots, p$가 주어졌다고 가정하자. 이 때, k번째 변수가 c_k개의 수준을 가진다면 행렬 $\mathbf{X}$의 k번째 열(column)은 크기 $n \times c_k$의 지시행렬(indicator matrix)

$$\mathbf{Z}_k = \left\{ z_{ih}^{(k)} \right\}, \quad i = 1, \ldots, n;\ h = 1, \ldots, c_k;\ k = 1, \ldots, p \tag{2.1}$$

로 표현할 수 있다. 이 때, 행렬 $\mathbf{Z}_k$의 원소 $z_{ih}^{(k)}$는 다음과 같이 정의된다.

$$z_{ih}^{(k)} = \begin{cases} 1, & i\text{번째 개체가 } k\text{번째 변수의 } h\text{번째 수준에 해당하는 경우,} \\ 0, & \text{그 외.} \end{cases}$$

따라서 주어진 자료행렬 $\mathbf{X}$는 다음과 같이 행렬 $\mathbf{Z}_k$를 부분행렬(submatrix)로 가지는 다중표시행렬(multiple indicator matrix) $\mathbf{Z}$로 재표현할 수 있다.

$$\mathbf{Z} = \left[\mathbf{Z}_1\ \mathbf{Z}_2 \cdots \mathbf{Z}_p \right]. \tag{2.2}$$

여기서, 행렬 $\mathbf{Z}$의 크기는 $n \times c$이며 $c = \sum_{k=1}^{p} c_k$이다. 그리고 식 (2.2)의 행렬 $\mathbf{Z}$을 이용하면 n개의 개체들의 비유사성은 단순매칭계수(simple matching coefficient)를 이용하여 측정할 수 있다. 그런데 이진수자료에 대한 단순매칭계수는 제곱유클리드거리(squared Euclidean distance)와 비례하므로, i번째 개체와 j번째 개체의 비유사성은

$$d_{ij}^2 = \sum_{k=1}^{p} \sum_{h=1}^{c_k} \left(z_{ih}^{(k)} - z_{jh}^{(k)} \right)^2$$

와 같이 정의 가능하며, 이를 이용하여 크기 $n \times n$의 비유사성행렬 $\mathbf{D}$를 다음과 같이 정의할 수 있다.

$$\mathbf{D} = \left\{ d_{ij}^2 \right\}, \quad i, j = 1, \ldots, n. \tag{2.3}$$

다음으로 Torgerson (1958)의 알고리즘을 이용하면, 비유사성행렬 $\mathbf{D}$의 정보를 다차원척도그림으로 표현할 수 있다. Young과 Householder (1938)를 바탕으로 한 Torgerson (1958)의 알고리즘은 계량형 다차원척도법의 대표적인 알고리즘으로 대수적으로 스펙트럼분해(spectral decomposition)를 통해 차원 축소된 형상공간의 좌표를 제공한다. 이에 비유사성행렬 $\mathbf{D}$를 이용한 계량형 다차원척도법의 적용 과정을 간략히 정리하면 다음과 같다.

1단계. 비유사성행렬 $\mathbf{D} = \{d_{ij}^2\}$로부터 행렬 $\mathbf{A}$를 다음과 같이 정의한다.

$$\mathbf{A} = -\frac{1}{2}\mathbf{D}.$$

2단계. 행렬 $\mathbf{A}$의 이중 중심화행렬 $\mathbf{B}$를

$$\mathbf{B} = \mathbf{HAH}$$

와 같이 구한다. 여기서 $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{J}_n$이고, $\mathbf{I}_n$은 n차 항등행렬(identity matrix), $\mathbf{J}_n$은 모든 원소가 1인 크기 $n \times n$의 행렬이다.

3단계. 다음과 같이 행렬 $\mathbf{B}$를 스펙트럼분해한다.

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^t.$$

여기서 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$으로 $\lambda_1 \geq \cdots \geq \lambda_n$의 관계를 만족하며, $\mathbf{V}\mathbf{V}^t = \mathbf{V}^t\mathbf{V} = \mathbf{I}_n$이다.

4단계. $\mathbf{V}_s$를 행렬 $\mathbf{V}$의 처음 s개 열로 이루어진 크기 $n \times s$의 행렬로 정의하고 $\mathbf{\Lambda}_s = \mathrm{diag}(\lambda_1, \ldots, \lambda_s)$라고 정의하면, n개 개체들에 대한 s차원 형상공간 상의 좌표(coordinates)를 나타내는 크기 $n \times s$의 행렬 $\mathbf{C}_s$는 다음과 같이 구할 수 있다.

$$\mathbf{C}_s = \mathbf{V}_s\mathbf{\Lambda}_s^{\frac{1}{2}}. \tag{2.4}$$

더불어 Gabriel (1971)에 따르면, 식 (2.4)의 행렬 $\mathbf{C}_s$에 의해 표현되는 s차원 다차원척도그림의 근사적 합도(goodness-of-fit of the approximation)는 다음과 같이 측정할 수 있다.

$$1 - \frac{||\mathbf{B} - \mathbf{B}_s||^2}{||\mathbf{B}||^2} = \frac{\sum_{i=1}^{s} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}, \tag{2.5}$$

여기서 $\mathbf{B}_s = \mathbf{C}_s\mathbf{C}_s^t$이고, $||\,\mathbf{B}\,|| = \sqrt{\mathrm{tr}(\mathbf{B}^t\mathbf{B})}$이다.

2.2. 분할법에 의한 범주형변수의 수준 정보 추가

식 (2.4)의 행렬 $\mathbf{C}_s$에 의해 표현되는 s차원 다차원척도그림은 개체들에 대한 정보만을 표현할 뿐 변수에 대한 정보는 표현되지 않아 변수들의 관점에서 개체들의 특성을 파악할 수 없다는 단점이 있다. Gower와 Harding (1988)은 계량형 다차원척도법이 가지는 이러한 단점을 보완하기 위하여 계량형자료에 대해 다차원척도법을 적용한 후, 개별 변수들이 가질 수 있는 변수값들에 대한 가상점들을 생성하고 이들 가상점들의 중심위치를 파악하여 다차원척도법의 형상공간에 투영하여 시각화하는 비선형행렬도를 제안한 바 있다. 비선형행렬도라 불리는 이유는 다차원척도법의 형상공간에 추가적으로 표현되는 계량형변수의 정보는 변수값이 변함에 따라 선형 혹은 비선형의 궤적을 남기기 때문이다. 이러한 비선

형행렬도의 기법을 계량형변수와 범주형변수를 모두 포함한 일반적인 자료에 대해 적용한 것이 Gower (1992)의 일반화 행렬도이다. 그리고 Shin 등 (2015)은 Gower (1992)의 방법을 응용하여 이진수자료에 대해 자료를 분할하여 각 변수의 2개 수준별 중심을 쉽고 빠르게 계산하고, 이들을 다차원척도법의 형상공간에 투영하는 분할법을 제안한 바 있다. 이에 본 연구에서는 Shin 등 (2015)의 분할법을 적용하여 식 (2.4)의 행렬 $\mathbf{C}_s$에 의해 표현되는 s차원 다차원척도그림에 각각의 범주수준별 정보를 추가적으로 시각화하는 방법을 제안하고자 한다.

분할법에 따르면, k번째 범주형변수가 가지는 c_k개의 수준별 중심은 식 (2.1)의 행렬 $\mathbf{Z}_k$와 식 (2.2)의 행렬 $\mathbf{Z}$를 이용하면 쉽게 계산할 수 있다. 우선, $k = 1, \ldots, p$에 대해 크기 $c_k \times c$의 행렬 $\mathbf{Y}_k$를 다음과 같이 정의하자.

$$\mathbf{Y}_k = \left(\mathbf{Z}_k^t \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^t \mathbf{Z}. \tag{2.6}$$

그리고 식 (2.6)의 행렬 $\mathbf{Y}_k$의 h번째 행을 크기 $c \times 1$의 벡터 $\mathbf{y}_h^{(k)}$, $h = 1, \ldots, c_k$, $k = 1, \ldots, p$라고 하면, 벡터 $\mathbf{y}_h^{(k)}$는 k번째 범주형변수가 h번째 수준을 가질 때 나머지 $p-1$개 범주형변수들의 각 수준별 발생비율 즉, k번째 범주형변수의 h번째 수준에 대한 조건부 확률분포(conditional probability distribution)를 의미하게 된다. 이러한 벡터 $\mathbf{y}_h^{(k)}$를 k번째 범주형변수의 h번째 수준의 중심이라 정의할 수 있다.

다음으로 벡터 $\mathbf{y}_h^{(k)}$의 정보를 식 (2.4)의 행렬 $\mathbf{C}_s$에 의해 표현되는 s차원 다차원척도법의 형상공간에 추가하기 위해서는 다음과 같은 과정이 필요하다. 우선, 행렬 $\mathbf{C}_s$에 의해 표현되는 s차원의 형상공간은 n개 개체들간의 비유사성 즉, 거리에 기반을 둔 공간이므로 k번째 범주형변수의 h번째 수준의 중심 $\mathbf{y}_h^{(k)}$와 관측된 n개 개체들간의 거리를 파악하여야 한다. 따라서 식 (2.2)의 행렬 $\mathbf{Z}$의 i번째 행을 벡터 $\mathbf{z}_i$, $i = 1, \ldots, n$라고 정의하면, i번째 관측된 개체와 k번째 범주형변수의 h번째 수준의 중심 $\mathbf{y}_h^{(k)}$ 사이의 거리는 식 (2.3)의 행렬 $\mathbf{D}$의 정의에 의해

$$a_{ih}^{(k)} = \left(\mathbf{z}_i - \mathbf{y}_h^{(k)}\right)^t \left(\mathbf{z}_i - \mathbf{y}_h^{(k)}\right), \quad i = 1, \ldots, n;\ h = 1, \ldots, c_k;\ k = 1, \ldots, p$$

와 같으며, k번째 범주형변수의 h번째 수준의 중심 $\mathbf{y}_h^{(k)}$에 대한 비유사성 벡터는 다음과 같이 크기 $n \times 1$의 벡터 $\mathbf{a}_h^{(k)}$로 정의할 수 있다.

$$\mathbf{a}_h^{(k)} = \left(a_{1h}^{(k)}, \ldots, a_{nh}^{(k)}\right)^t, \quad h = 1, \ldots, c_k;\ k = 1, \ldots, p. \tag{2.7}$$

여기서, 비유사성에 기반을 둔 s차원 다차원척도법의 형상공간에 k번째 범주형변수의 h번째 수준의 중심 $\mathbf{y}_h^{(k)}$의 정보를 추가한다는 것은 식 (2.4)의 행렬 $\mathbf{C}_s$의 벡터공간에 식 (2.7)의 비유사성 벡터 $\mathbf{a}_h^{(k)}$를 투영하는 것을 의미한다. 따라서 Torgerson (1958)의 알고리즘의 1단계와 2단계에 맞춰 중심화된 벡터 $\mathbf{a}_h^{(k)}$를 행렬 $\mathbf{C}_s$의 벡터공간에 투영한 k번째 범주형변수의 h번째 수준에 대한 좌표벡터 $\mathbf{c}_h^{(k)}$는 다음과 같이 정의된다.

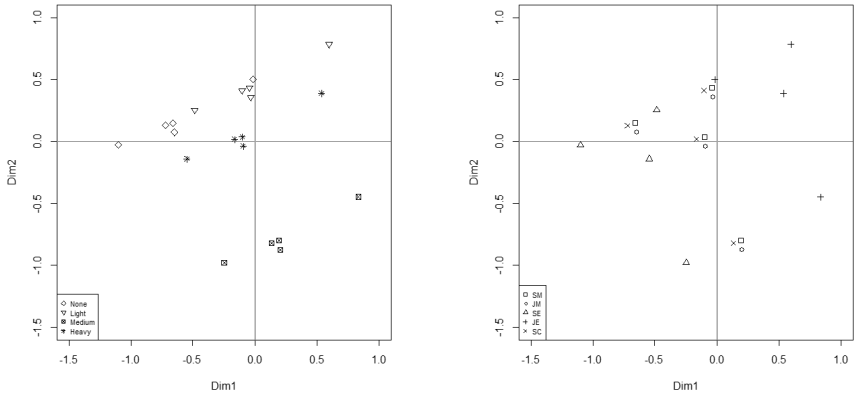
$$\mathbf{c}_h^{(k)} = \mathbf{\Lambda}_s^{-1} \mathbf{C}_s^t \left(-\frac{1}{2}\mathbf{a}_h^{(k)} - \frac{1}{n}\mathbf{A}\mathbf{1}_n\right), \quad h = 1, \ldots, c_k;\ k = 1, \ldots, p. \tag{2.8}$$

이 때, $\mathbf{1}_n$은 모든 원소가 1인 크기 $n \times 1$ 벡터이다. 식 (2.8)에 의해 계산된 $\mathbf{c}_h^{(k)}$는 크기 $s \times 1$의 벡터이며, 이를 2.1절에서 생성한 s차원 다차원척도그림에 추가적으로 표현하면 k번째 범주형변수의 h번째 수준의 정보를 파악하는데 도움을 얻을 수 있다.

분할법에 의한 범주형변수의 수준 정보를 추가하는 과정을 간략히 정리하면 다음과 같다.

Table 3.1. Smoking habits according to position

Positions	Smoking habits				Total
	None	Light	Medium	Heavy	
SM	4	2	3	2	11
JM	4	3	7	4	18
SE	25	10	12	4	51
JE	18	24	33	13	88
SC	10	6	7	2	25



(a) Smoking habits (b) Positions

Figure 3.1. Multidimensional scaling maps of positions and smoking habits.

1단계. 식 (2.6)에 의해 k번째 범주형변수가 가지는 c_k개의 수준별 중심을 행으로 가지는 크기 $c_k \times c$의 행렬 $\mathbf{Y}_k$, $k = 1, \ldots, p$를 계산한다.

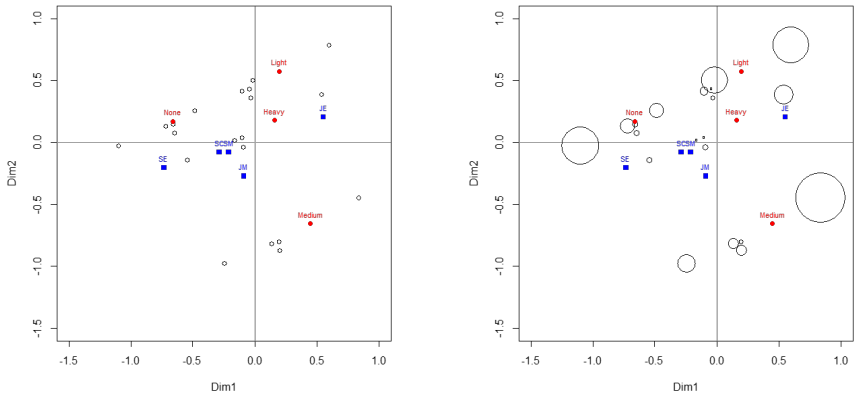
2단계. 행렬 $\mathbf{Y}_k$의 각각의 행 $\mathbf{y}_h^{(k)}$, $h = 1, \ldots, c_k$와 i번째 관측된 개체 $\mathbf{z}_i$, $i = 1, \ldots, n$의 거리를 파악하여 식 (2.7)의 k번째 범주형변수의 h번째 수준의 중심 $\mathbf{y}_h^{(k)}$에 대한 비유사성 벡터 $\mathbf{a}_h^{(k)}$를 구한다.

3단계. 식 (2.8)에 의해 벡터 $\mathbf{a}_h^{(k)}$를 행렬 $\mathbf{C}_s$의 $s(< p)$차원 벡터공간에 투영한 k번째 범주형변수의 h번째 수준에 대한 중심좌표 $\mathbf{c}_h^{(k)}$를 구하고, 이를 s차원 다차원척도그림 상에 추가한다.

3. 활용 사례

Table 3.1은 어느 회사에서 직원 193명을 대상으로 조사한 직위에 따른 흡연습관을 나타낸 이원분할표이다 (Choi와 Shin, 2013, Chapter 3). 개인의 흡연 습관은 하루 한 개비도 피우지 않는 경우(none)와 하루에 10개비 이하로 피는 경우(light)로, 하루에 11개비에서 19개비를 피우는 경우(medium), 하루에 1갑 이상 피는 경우(heavy)의 4개 수준으로 구분되어 있으며, 직원들의 직위는 상위 경영직(SM), 하위 경영직(JM), 상위 고용직(SE), 하위 고용직(JE), 비서직(SC)의 5개 수준으로 구분되어 있다.

주어진 자료를 이용하여 2.1절에서 설명한 방법에 따라 193명 직원들의 비유사성을 측정하고 식 (2.4)의 행렬 $\mathbf{C}_s$에 의해 표현되는 2(= s)차원의 다차원척도그림이 Figure 3.1에 제시되어 있다. 실



(a) General map (b) Map with relative size of groups

Figure 3.2. Proposed multidimensional scaling maps.

Table 3.2. Coordinates of centroids for each level of positions and smoking habits

	SM	JM	SE	JE	SC	None	Light	Medium	Heavy
Dim1	−0.213	−0.092	−0.736	0.554	−0.289	−0.660	0.194	0.445	0.157
Dim2	−0.078	−0.272	−0.205	0.206	−0.077	0.172	0.573	−0.656	0.179

제로 Figure 3.1의 (a)와 (b)는 동일한 그림이다. 단지 다차원척도그림의 해석을 도모하고자 직위와 흡연습관의 범주수준별로 좌표점의 모양을 다르게 표현하였을 뿐이다. Figure 3.1에 제시되어 있는 2차원 다차원척도그림에 대해 식 (2.5)의 근사적합도를 계산한 결과는 66.73%로 주성분분석 등에서 활용되는 적합도의 기준인 70%에는 못 미치나 부족한 수준은 아니라고 판단된다.

더불어 Figure 3.1의 다차원척도그림에 좌표점의 개수가 20개만 표시되는 이유는 주어진 자료가 범주형자료이기에 동일한 관측값을 가지는 개체들이 많으므로, 두 변수의 수준조합 개수인 20(= 흡연습관(4수준) × 직위(5수준))개로 겹쳐져서 표현되기 때문이다. Figure 3.1의 (a)를 살펴보면, 우측 하단으로는 Medium에 해당하는 군집이 위치함을 확인할 수 있다. 또한, Heavy에 해당하는 군집은 원점 주변에 위치하며, None과 Light에 해당하는 군집은 각각 좌측과 상단에 위치함을 알 수 있다. 그리고 Figure 3.1의 (b)로부터 SE는 좌측 하단으로 산재되어 있으며, JE은 우측 상단으로 산재되어 있으므로 타직급과는 이질적임을 확인할 수 있다.

그러나 Figure 3.1의 다차원척도그림만으로는 개체들의 군집화 성향만을 확인할 수 있을 뿐, 군집들의 특징을 파악하거나 범주수준들 사이의 연관성을 파악하기에는 어려움이 있다. 따라서 식 (2.8)에 의해 2개의 범주형변수 즉, 흡연습과 직위에 대한 각 수준별 중심을 다차원척도그림에 투영한 좌표가 Table 3.2에 제시되어 있으며, Figure 3.2는 이들을 표시한 다차원척도그림을 보여준다. 여기서 Figure 3.2의 (a)와 (b)는 동일한 그림으로, Figure 3.2의 (a)에서 개체들의 군집을 나타내는 20개 좌표점들에 대해 각 군집의 발생비율에 따라 좌표점의 크기를 재조정하면 Figure 3.2의 (b)와 같이 표현된다. 즉, Figure 3.2의 (b)에서 개체들에 대한 20개 좌표점들의 크기는 전체 193명에 대한 20개 수준조합의 상대도수에 비례하게 표현되어 있으므로, 좌표점의 크기가 클수록 해당 군집에 속하는 개체수가 많음을 확인할 수 있을 뿐만 아니라 주어진 범주형변수들의 결합확률분포(joint probability distribution)를 파악할 수 있다.

Table 3.3. Centroids for each level of positions and smoking habits

		SM	JM	SE	JE	SC	None	Light	Medium	Heavy
$\mathbf{Y}_1$	SM	1.000	0.000	0.000	0.000	0.000	0.364	0.182	0.273	0.182
	JM	0.000	1.000	0.000	0.000	0.000	0.222	0.167	0.389	0.222
	SE	0.000	0.000	1.000	0.000	0.000	0.490	0.196	0.235	0.078
	JE	0.000	0.000	0.000	1.000	0.000	0.205	0.273	0.375	0.148
	SC	0.000	0.000	0.000	0.000	1.000	0.400	0.240	0.280	0.080
$\mathbf{Y}_2$	None	0.066	0.066	0.410	0.295	0.164	1.000	0.000	0.000	0.000
	Light	0.044	0.067	0.222	0.533	0.133	0.000	1.000	0.000	0.000
	Medium	0.048	0.113	0.194	0.532	0.113	0.000	0.000	1.000	0.000
	Heavy	0.080	0.160	0.160	0.520	0.080	0.000	0.000	0.000	1.000

또한, Table 3.3에는 식 (2.6)에 의해 계산된 2개의 행렬 $\mathbf{Y}_1$과 $\mathbf{Y}_2$가 제시되어 있다. 이들의 각각의 행은 2.2절에서 각 수준별 조건부 확률분포를 나타낸다는 것을 언급한 바 있다. 예를 들어, 행렬 $\mathbf{Y}_1$의 첫 번째 행 SM의 경우 SM 수준의 발생 조건이 주어졌기에 SM의 발생비율은 1인 반면, JM, SE, JE, SC의 발생비율은 0으로 나타났다. 또한 Table 3.1로부터 SM의 경우 None에 해당하는 비율은 $4/11 \doteq 0.364$이고 Light, Mdeium, Heavy에 해당하는 비율은 각각 $0.182(\doteq 2/11)$, $0.273(\doteq 3/11)$, $0.182(\doteq 2/11)$임을 확인할 수 있다. 더불어 Figure 3.2에 표현되어 있는 수준별 중심들의 좌표는 이들 각 범주형 변수들의 수준별 조건부 확률분포가 반영되어 있으므로 각각의 중심들의 성향을 해석하는데 참고할 수 있다.

다차원척도법에 의한 형상공간은 실제 개체간의 비유사성을 근사적으로 표현하는 공간이므로, 다차원척도그림 상에 추가된 수준별 중심과 개체들의 군집을 나타내는 좌표점이 가까울수록 해당 수준의 성향이 높다고 해석할 수 있다. 따라서 Figure 3.2를 살펴보면, 우측 하단으로는 흡연습관 중 Medium 수준의 중심이 위치하고 있다. 따라서 해당 좌표점 주변의 개체 군집들은 하루에 11개비에서 19개비의 담배를 피우는 군집이라고 할 수 있다. 그리고 Figure 3.2의 가장 좌측에 위치한 개체들의 군집은 흡연습관은 None과 가까우며, 직위는 SE와 가까움을 확인할 수 있다. 따라서 해당 군집은 담배를 피지않는 상위고용직이라고 유추할 수 있다. 더불어 SE는 None을 제외한 나머지 수준의 흡연습관과는 상대적으로 멀리 위치함으로써 상위고용직에는 상대적으로 비흡연자가 많음을 유추할 수 있다. Table 3.3의 3번째 행으로부터 상위고용직의 비흡연자 비율이 0.490으로 상대적으로 높음을 확인할 수 있다. 다음으로 Figure 3.2의 가장 우측에는 직위 중 JE 수준의 중심이 위치하는데, 해당 좌표는 흡연 중임을 나타내는 3개 흡연습관의 중심과 가까이 위치함을 확인할 수 있다. 이를 통해 하위고용직 종사자들은 상대적으로 흡연자가 많음을 유추할 수 있으며, Table 3.3의 4번째 행으로부터 하위고용직의 흡연자 비율이 0.795로 상당히 높음을 확인할 수 있다. 특히, Table 3.3의 9번째 행으로부터 Heavy 즉, 하루에 1갑 이상의 담배를 피는 직원들 중에는 하위고용직 종사자의 비율이 0.520로 상대적으로 비율이 높음을 확인할 수 있는데, 이는 Figure 3.2에서 JE의 좌표점이 Heavy의 좌표점과 가깝게 위치하는 것을 통해 파악 가능하다.

4. 결론

다차원척도법은 저차원 공간에 개체간의 비유사성을 최대한 유사하게 나타내기 위한 다변량 자료의 탐색적 분석기법이다. 특히, 계량형 다차원척도법은 일반적으로 계량형 자료를 이용하여 개체 간의 거리를 표현함으로써 개체들의 군집화 성향을 탐색하기 위해 많이 활용된다. 그러나 일반적인 다차원척도법에 의해 표현되는 다차원척도그림에서는 변수와 관련된 정보가 나타나지 않기 때문에 그림의 해석 상에

한계점이 존재한다.

본 연구에서는 범주형자료를 다중표시행렬로 변환하고 Torgerson (1958)의 알고리즘을 이용하여 개체들의 군집화 성향을 저차원의 다차원척도그림에 표현하였다. 그리고 Shin 등 (2015)의 분할법을 적용하여 범주형변수의 수준별 정보를 다차원척도 그림 상에 추가함으로써 범주형자료를 쉽고 빠르게 탐색할 수 있는 시각화 방법을 제안하였다. 본 연구에서 제안된 다차원척도그림에서는 개체군집들에 대한 좌표점의 크기를 통해 변수들의 결합확률분포에 대한 탐색이 가능하며, 좌표점의 위치를 통해 개체들의 군집화 성향을 탐색할 수 있는 장점이 있다. 또한, 범주형변수들의 수준별 중심에 대한 좌표는 수준별 중심위치의 계산 방법이 조건부 확률분포와 관련이 있으므로, 다차원척도그림 상에 추가된 수준별 중심좌표의 위치를 통해 범주형변수들 사이의 연관성을 탐색할 수 있는 장점도 있다. 그러나 다른 유사성측도를 적용하는 경우와의 비교를 포함한 추가적인 연구가 필요하다고 생각된다.

References

Choi, Y. S. (2014). *Walk in Multidimensional Scaling*, Free Academy, Seoul.

Choi, Y. S. and Shin, S. M. (2013). *Understanding of Biplot Analysis using R*, Free Academy, Seoul.

Cox, T. F. and Cox, M. A. A. (2000). *Multidimensional Scaling* (2nd ed), Chapman & Hall, London.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453–467.

Gower, J. C. (1992). Generalized biplots, *Biometrika*, **79**, 475–493.

Gower, J. C. and Hand, D. J. (1996). *Biplots*, Chapman & Hall, London.

Gower, J. C. and Harding, S. A. (1988). Nonlinear biplots, *Biometrika*, **75**, 445–455.

Shin, S. M., Kim, E. S., and Choi, Y. S. (2015). Multidimensional scaling using the pseudo-points based on partition method, *Korean Journal of Applied Statistics*, **28**, 1171–1180.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*, Wiley, New York.

Young, G. and Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances, *Psychometrika*, **3**, 19–22.

분할법을 활용한 범주형자료의 다차원척도법

신상민[a] · 천선경[b] · 최용석[b,1]

[a]동아대학교 경영정보학과, [b]부산대학교 통계학과

(2017년 10월 23일 접수, 2017년 12월 5일 수정, 2017년 12월 13일 채택)

요 약

다차원척도법은 개체간의 비유사성을 저차원 공간에 기하적으로 표현하기 위한 다변량 자료의 탐색적 분석기법이다. 그러나 일반적인 다차원척도그림에서는 개체들의 유사성 정보만이 표현될 뿐 변수와 관련된 정보가 나타나지 않기 때문에 그림의 해석 상에 한계점이 존재한다. 본 연구에서는 범주형 자료를 다중표시행렬로 변환하고 Torgerson (1958)의 알고리즘에 의한 다차원척도법을 적용하여 개체들의 군집화 성향과 군집들의 상대적 크기를 다차원척도그림으로 시각화하였다. 그리고 Shin 등 (2015)의 분할법을 적용하여 범주형변수의 범주수준별 정보를 다차원척도그림 상에 투영하여 추가적인 정보를 표현하였다. 따라서 본 연구에서 제안하고자 하는 다차원척도그림을 이용하면 개체들의 유사성 정보와 함께 범주형변수들 사이의 연관성도 탐색할 수 있는 장점이 있다.

주요용어: 다차원척도법, 범주형자료, 분할법, 시각화

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

[1]교신저자: (46241) 부산광역시 금정구 부산대학로63번길 2, 부산대학교 통계학과. E-mail: yschoi@pusan.ac.kr