The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1171
28(6), 1171–1180

Multidimensional Scaling Using the Pseudo-Points Based on Partition Method

Sang Min Shin[a] · Eun-Seong Kim[a] · Yong-Seok Choi[a,1]

[a]Department of Statistics, Pusan National University

(Received October 12, 2015; Revised December 1, 2015; Accepted December 14, 2015)

Abstract

Multidimensional scaling (MDS) is a graphical technique of multivariate analysis to display dissimilarities among individuals into low-dimensional space. We often have two kinds of MDS which are metric MDS and non-metric MDS. Metric MDS can be applied to quantitative data; however, we need additional information about variables because it only shows relationships among individuals. Gower (1992) proposed a method that can represent variable information using trajectories of the pseudo-points for quantitative variables on the metric MDS space. We will call his method a 'replacement method'. However, the trajectory can not be represented even though metric MDS can be applied to binary data when we apply his method to binary data. Therefore, we propose a method to represent information of binary variables using pseudo-points called a 'partition method'. The proposed method partitions pseudo-points, accounting both the rate of zeroes and ones. Our metric MDS using the proposed partition method can show the relationship between individuals and variables for binary data.

Keywords: multidimensional scaling, pseudo-points, replacement method, partition method

1. 서론

다차원척도법(MDS)이란, 다차원 공간에서 정의되는 유사성(similarity) 또는 비유사성(dissimilarity)을 저차원의 공간에 기하적으로 나타내어 그들의 관계를 탐색적으로 살펴보는 다변량 그래프적 기법이다. 특히 다차원척도법에서 표현되는 저차원 공간을 형상공간(configuration space)이라 하며, 여기에 개체들의 정보를 기하적으로 나타낸 것을 다차원척도법도(MDS map)라 한다 (Choi, 2014). 이러한 다차원척도법은 저차원 공간상에 개체들의 유사성 정보만을 나타낼 뿐 변수들의 정보는 나타내지 못한다는 단점이 있다. 이에 Gower (1968)는 다차원척도법의 형상공간상에 가상점을 활용하여 변수들의 정보를 나타내는 축을 추가시킴으로서 개체와 변수간의 관계를 파악할 수 있는 비선형 행렬도(nonlinear biplot)를 제안하였다. 비선형 행렬도는 양적 변수에 의해 측정된 개체들의 유사성을 표현한 다차원척도법도의 공간 상에 일반화시킨 변수들의 정보를 가상점들(pseudo-points)로 표현해내고, 이들 가상점의 궤적(trajectory)을 파악하여 변수정보를 표현하는 축을 추가하게 된다. 그러나 이러한 비선형 행렬도는 양적 변수에 의해 측정된 자료에 대해서만 적용할 수 있는 기법이기에 Gower (1992)와 Gower와

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

Hand (1996)는 양적 변수와 범주형 변수가 모두 포함된 자료의 경우, 일반화 행렬도(generalized biplots)를 사용할 것을 제안하였다. 일반화 행렬도는 비선형 행렬도와 유사하게 개체들의 유사성 정보를 다차원척도법도의 공간 상에 표현한 후, 가상점을 활용하여 양적 변수들의 정보를 나타내는 축과 범주형 변수들의 정보를 나타내는 점을 추가하게 된다. 이러한 일반화 행렬도에서는 각각의 변수들에 대한 가상점을 표현하기 위해 대체법(replacement method)을 사용하는데, 이는 주어진 자료행렬에서 특정 변수의 관측값을 임의의 값으로 대체한 가상점을 정의하고 이들 가상점의 중심을 다차원척도법의 형상공간에 투영(projection)하여 변수 정보를 표현하는 방식이다.

0과 1 두 개의 관측값만을 갖는 이진수 자료의 경우, 자료의 특성 상 양적 변수의 경우와 같은 방법으로 다차원척도법을 적용할 수 있다. 그러나 변수 정보를 표현하기 위해 기존의 대체법을 적용하면 0과 1의 두 범주 각각의 성향을 파악하기 어려우므로, 본 연구에서는 개별 변수의 변수값이 0인 가상점과 1인 가상점을 분할하여 고려하는 분할법(partition method)을 제안하고자 한다. 이에 2절에서는 계량형 다차원척도법의 이진수 자료에 대한 적용을 설명하고 더불어 대체법과 분할법을 이용하여 다차원척도법 공간상에 가상점을 표현하는 방법을 설명한 후, 3절에서는 분할법의 활용 사례를 통해 대체법의 적용 결과와 비교하려한다. 끝으로 4절에서 본 연구를 정리 및 요약하려 한다.

2. 가상점을 활용한 다차원척도법

2.1. 다차원척도법

이 절에서는 Choi (2014)과 Torgerson (1958)을 참고로 하여 이진수 자료에 계량형 다차원척도법의 적용을 소개하며 기초 이론을 요약하고자 한다. 일반적으로 다차원척도법은 i번째 개체와 j번째 개체간의 비유사성 d_{ij}와 차원축소된 형상공간에서 거리 δ_{ij} 사이의 관계가 일치되도록 표현하는 자료 축약 기법으로, 이들 d_{ij}와 δ_{ij}의 관계는 다음과 같이 모형화 할 수 있다.

$$d_{ij} = f(\delta_{ij}) + \varepsilon, \quad i, j = 1, \ldots, n, \tag{2.1}$$

여기서 ε_{ij}는 측정 및 저차원 공간 근사에 따른 왜곡오차를 의미하고, $f(\cdot)$는 단조함수로 정의된다. Everitt와 Dunn (1991)과 Huh (1994)에 따르면, 식 (2.1)에서 상수 a와 b에 대해 $d_{ij} = \delta_{ij} + \varepsilon_{ij}$이면 절대척도라 하고 $d_{ij} = a + b\delta_{ij} + \varepsilon_{ij}$이면 구간척도, $d_{ij} = b\delta_{ij} + \varepsilon_{ij}$, $b > 0$이면 비율척도라 하며, 모든 $i, j, k, l = 1, \ldots, n$에 대하여 $d_{ij} \leq d_{kl}$이면 $f(d_{ij}) \leq f(d_{kl})$가 성립하면 순서척도라 한다. 특히, Kruskal과 Wish (1978)는 d_{ij}와 δ_{ij}의 관계가 절대척도나 구간척도, 비율척도를 만족하는 다차원척도법을 계량형 다차원척도법(metric MDS)이라 하였고, 순서척도를 만족하는 경우는 비계량형 다차원척도법(nonmetric MDS)이라 하였다.

p개의 이진수 변수에 대해 n개의 개체를 측정한 자료행렬 $\mathbf{X} = (x_{ik})$, $i = 1, \ldots, n$; $k = 1, \ldots, p$를

$$x_{ik} = \begin{cases} 1, & i\text{번째 개체가 } k\text{번째 변수의 성질을 만족하는 경우,} \\ 0, & \text{그 외의 경우} \end{cases} \tag{2.2}$$

와 같이 정의하면, i번째 개체 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t$와 j번째 개체 $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})^t$ 사이의 비유사성은 보편적으로 Mardia 등 (1979)과 Cox와 Cox (1994)에 따라 단순매칭계수를 이용하여 측정 가능하다. 그런데 이러한 단순매칭계수에 의한 비유사성은 양적자료의 보편적 비유사성 측정 방식인 두 개체 사이의 제곱유클리드거리

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^t(\mathbf{x}_i - \mathbf{x}_j), \quad i, j = 1, \ldots, n \tag{2.3}$$

를 p로 나눈 것과 일치하게 된다. 따라서 일반적으로 계량형 다차원척도법은 양적자료에 적용하는 기법임에도 불구하고, 질적자료인 이진수 자료는 특별히 계량형 다차원척도법의 적용이 가능하다.

계량형 다차원척도법의 대표적인 알고리즘인 토거선 알고리즘 (Torgerson, 1958)은 비유사성 행렬에 대한 스펙트럼분해(spectral decomposition)를 통해 차원 축소된 형상공간의 좌표를 제공한다. 이러한 토거선 알고리즘에 대해 간략히 정리하면 다음과 같다. 우선, 비유사성행렬 $\mathbf{D} = (d_{ij}^2)$로 부터 행렬 $\mathbf{A}$를 정의하는데, 여기서 행렬 $\mathbf{A}$의 원소 a_{ij}는 다음과 같다.

$$a_{ij} = -\frac{1}{2}d_{ij}^2, \quad i, j = 1, \ldots, n. \tag{2.4}$$

다음으로 행렬 $\mathbf{A}$로부터 이중-중심화행렬

$$\mathbf{B} = (b_{ij}) = \mathbf{HAH}$$

를 계산한다. 여기서, $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{J}_n$이고 $\mathbf{I}_n$은 n차 항등행렬, $\mathbf{J}_n$는 모든 원소가 1인 n차 정방행렬을 의미한다. 즉, 행렬 $\mathbf{A}$의 i번째 행의 평균을 $\bar{a}_{i\cdot} = \sum_{j=1}^{n} a_{ij}/n$이라 하고, j번째 열의 평균을 $\bar{a}_{\cdot j} = \sum_{i=1}^{n} a_{ij}/n$, 모든 원소의 평균을 $\bar{a}_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}/n^2$이라 하면, 행렬 $\mathbf{B}$의 원소 b_{ij}는 다음과 같다.

$$b_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$$

다음 단계에서는 차원 축소된 형상공간의 좌표를 얻기 위해 행렬 $\mathbf{B}$의 스펙트럼분해

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

를 이용한다. 여기서 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$는 $\lambda_1 \geq \cdots \geq \lambda_n$의 관계를 갖는 고유값을 대각원소로 갖는 대각행렬이며, $\mathbf{V}$는 고유벡터 $\mathbf{v}_1, \ldots, \mathbf{v}_n$를 열로 가지는 크기 $n \times n$의 직교행렬이다. 이러한 행렬 $\mathbf{B}$의 스펙트럼분해로부터 처음 $s\,(\leq p)$개의 고유값과 이에 대응하는 고유벡터를 이용하여 크기가 $n \times s$인 행렬 $\mathbf{C}_{(s)}$를 다음과 같이 정의할 수 있다.

$$\mathbf{C}_{(s)} = \mathbf{V}_{(s)}\Lambda_{(s)}^{\frac{1}{2}} = (\mathbf{v}_1\lambda_1, \ldots, \mathbf{v}_s\lambda_s). \tag{2.5}$$

식 (2.5)의 행렬 $\mathbf{C}_{(s)}$가 s차원 다차원척도법도의 n개 개체들에 대한 좌표점을 제공한다. 일반적으로, s차원 다차원척도법도의 근사적합도(goodness-of-fit of the approximation)는

$$\frac{\sum_{r=1}^{s}\lambda_r}{\sum_{r=1}^{n}\lambda_r} \times 100\% \tag{2.6}$$

와 같으며, 일반적으로 식 (2.6)의 근사적합도는 다차원척도법도의 설명력을 나타낸다.

2.2. 가상점 표현

2.2.1. 대체법에 의한 가상점 표현 앞서 언급한 바와 같이 다차원척도법은 저차원 공간상에 개체들의 유사성 정보만을 나타낼 뿐 변수들의 정보는 나타내지 못한다는 단점이 있다. 이에 이 절에서는 Choi와 Shin (2013)과 Gower (1992), Gower와 Hand (1996)를 참고하여 일반화 행렬도에서 변수 정보를 표현하기 위해 사용하는 대체법에 대해 설명하고자 한다.

먼저 p개의 양적 변수에 대해 n개의 개체를 측정한 자료 행렬을 $\mathbf{X} = (x_{ik})$, $i = 1, \ldots, n$; $k = 1, \ldots, p$라 하자. 그러면 k번째 변수에 대한 정보를 얻기 위해서 주어진 자료 행렬 $\mathbf{X}$의 k번째 열의 관측값을 해당 변수가 가질 수 있는 특정한 값 τ로 대체한 행렬 $\mathbf{X}^* = (x_{ik}^*)$을 다음과 같이 정의할 수 있다.

$$\mathbf{X}^* = \begin{pmatrix} x_{11}^* & \cdots & x_{1;k-1}^* & x_{1;k}^* & x_{1;k+1}^* & \cdots & x_{1p}^* \\ & & & \vdots & & & \\ x_{i1}^* & \cdots & x_{i;k-1}^* & x_{i;k}^* & x_{i;k+1}^* & \cdots & x_{ip}^* \\ & & & \vdots & & & \\ x_{n1}^* & \cdots & x_{n;k-1}^* & x_{n;k}^* & x_{n;k+1}^* & \cdots & x_{np}^* \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1;k-1} & \tau & x_{1;k+1} & \cdots & x_{1p} \\ & & & \vdots & & & \\ x_{i1} & \cdots & x_{i;k-1} & \tau & x_{i;k+1} & \cdots & x_{ip} \\ & & & \vdots & & & \\ x_{n1} & \cdots & x_{n;k-1} & \tau & x_{n;k+1} & \cdots & x_{np} \end{pmatrix}.$$

이러한 행렬 $\mathbf{X}^*$의 각각의 행을 가상점이라 하며, τ의 값이 변함에 따른 이들 n개 가상점의 중심에 대한 다차원척도법도 상에서의 궤적이 해당 변수의 정보를 표현해 주는 축이 된다.

다차원척도법도의 형상공간은 개체간의 거리에 기반을 둔 공간이므로 가상점의 중심을 기존 개체들에 의한 형상공간에 투영하기 위해서는 기존의 n개 좌표점과 추가로 정의된 n개 가상점과의 거리가 반영되어야 한다. 따라서 기존의 i번째 개체 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik}, \ldots, x_{ip})^t$와 j번째 가상점 $\mathbf{x}_j^* = (x_{j1}, \ldots, \tau, \ldots, x_{jp})^t$ 간의 변환된 제곱유클리드거리를

$$m_{ij}^2 = -\frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j^*\right)^t \left(\mathbf{x}_i - \mathbf{x}_j^*\right), \quad i, j = 1, \ldots, n$$

라 하고, 이들 거리에 의한 비유사성 행렬을 $\mathbf{M} = (m_{ij}^2)$이라 하자. 다음으로 i번째 가상점 $\mathbf{x}_i^* = (x_{i1}, \ldots, \tau, \ldots, x_{ip})^t$와 j번째 가상점 $\mathbf{x}_j^* = (x_{j1}, \ldots, \tau, \ldots, x_{jp})^t$간의 변환된 제곱유클리드거리를

$$p_{ij}^2 = -\frac{1}{2}\left(\mathbf{x}_i^* - \mathbf{x}_j^*\right)^t \left(\mathbf{x}_i^* - \mathbf{x}_j^*\right), \quad i, j = 1, \ldots, n$$

라 하고, 이들 거리에 의한 비유사성 행렬을 $\mathbf{P} = (p_{ij}^2)$이라 하자.

Gower와 Hand (1996)는 행렬 $\mathbf{M}$과 $\mathbf{P}$를 이용하여 n개 가상점의 중심과 기존의 n개 좌표점과의 제곱거리를 나타내는 벡터를 다음과 같이 정의하였다.

$$\boldsymbol{a} = \frac{\mathbf{1}_n \mathbf{P} \mathbf{1}_n}{n^2} \mathbf{1}_n - \frac{n}{2} \mathbf{M} \mathbf{1}_n, \tag{2.7}$$

여기서 $\mathbf{1}_n$은 모든 원소가 1인 $n \times 1$ 벡터이다. 그런데 식 (2.7)의 연산은 가상점들에 의한 행렬 $\mathbf{X}^*$의 열평균

$$\bar{\mathbf{x}}^* = \left(\bar{x}_1^*, \ldots, \bar{x}_k^*, \ldots, \bar{x}_p^*\right)^t = (\bar{x}_1, \ldots, \bar{x}_{k-1}, \tau, \bar{x}_{k+1}, \ldots, \bar{x}_p)^t \tag{2.8}$$

과 기존의 n개 개체 사이의 제곱유클리드거리와 같다. 여기서, $\bar{x}_k^* = (1/n)\sum_{i=1}^n x_{ik}^*$이다. 이러한 이유로 식 (2.7)의 거리벡터 $\boldsymbol{a}$를 n개 가상점의 중심과 기존의 n개 좌표점과의 제곱거리를 나타내는 벡터라고 한다. 그러므로 식 (2.7)의 거리벡터 $\boldsymbol{a} = (a_1, \ldots, a_n)^t$의 원소는

$$a_i = (\mathbf{x}_i - \bar{\mathbf{x}}^*)^t (\mathbf{x}_i - \bar{\mathbf{x}}^*), \quad i = 1, \ldots, n \tag{2.9}$$

와 같이 재표현할 수 있으며, Gower (1968)는 이를 이용하여 가상점들의 중심을 s차원으로 축소된 다차원척도법의 형상공간에 투영한 좌표를 다음과 같이 구하였다.

$$\mathbf{c}_k = \mathbf{\Lambda}_{(s)}^{-1} \mathbf{C}_{(s)}^t \left[-\frac{1}{2}\boldsymbol{a} - \frac{1}{n}\mathbf{A}\mathbf{1}_n\right], \quad k = 1, \ldots, p. \tag{2.10}$$

따라서 τ의 값이 변함에 따라 생성되는 좌표 $\mathbf{c}_k$의 궤적이 k번째 변수의 정보를 표현해 주는 축이 된다.

2.2.2. 분할법에 의한 가상점 표현 앞서 2.1절에서 이진수 자료에 대해 계량형 다차원척도법을 적용할 수 있음을 언급한 바 있다. 따라서 Gower와 Hand (1996)의 대체법을 이진수 자료의 다차원척도법에 동일하게 적용하면, 식 (2.9)와 식 (2.10)를 이용하여 가상점의 중심에 대한 좌표는 구할 수 있다. 그러나 이진수 변수는 0과 1의 값만을 가질 수 있으므로, 해당 변수의 정보는 다차원척도법의 형상공간상에 궤적을 남기지 못하고 두 개의 점으로 표현된다. 더불어 대체법을 적용하면 식 (2.8)에 의해 가상점의 중심을 계산한 경우, k번째 변수를 제외한 변수들의 관측값 1의 비율에 따라 가상점의 중심이 결정되어, 본 연구에서 적용한 이진수 자료의 활용 사례에 대체법을 적용한 결과 가상점이 원점에 지나치게 가까이 표현되어 해석이 어려워지는 문제점을 발견하였다. 따라서 이진수 자료에 대해서는 주어진 자료행렬로부터 생성되는 가상점들을 k번째 변수에 해당하는 관측값이 0인 가상점과 1인 가상점 분할하여 고려한 후, 두 개의 분할된 가상점의 중심을 계산하는 분할법을 이용할 것을 제안하고자 한다.

분할법은 이진수 변수의 성향 파악 과정에서 k번째 변수가 각각 0과 1의 값을 가지는 경우의 조건부확률분포를 반영하기 위하여 크기 $n \times p$의 주어진 이진수 자료행렬 $\mathbf{X}$를 새로운 새로운 n개의 가상점에 의한 행렬 $\mathbf{X}^*$와 동일시하고, 이들 가상점을 두 개의 그룹으로 분할하여 그룹별 성향을 각각 파악하고자 한다. 이에 $\mathbf{X}^* = \mathbf{X}$이므로, 이진수 자료행렬 $\mathbf{X}$의 k번째 열의 정보를 재표현한 크기 $n \times 2$의 분할계획행렬(partition design matrix) $\mathbf{G}_k = (g_{ih}^{(k)})$, $i = 1, \ldots, n$; $k = 1, \ldots, p$; $h = 1, 2$

$$g_{ih}^{(k)} = \begin{cases} 1, & i\text{번째 개체가 } k\text{번째 변수에 대해 } h\text{번째 범주에 해당하는 경우,} \\ 0, & \text{기타} \end{cases}$$

를 정의하면, 이러한 행렬 $\mathbf{G}_k$의 내적인 행렬 $\mathbf{D}_k$는

$$\mathbf{D}_k = \mathbf{G}_k^t \mathbf{G}_k \begin{bmatrix} \sum_{i=1}^{n} g_{i1}^{(k)} & 0 \\ 0 & \sum_{i=1}^{n} g_{i2}^{(k)} \end{bmatrix} \tag{2.11}$$

와 같이 k번째 변수의 h번째 범주에 해당하는 개체의 수를 대각원소로 가지는 대각행렬이 된다. 이는 해당 범주에 속하는 가상점의 수와 같으므로, k번째 변수의 값이 0인 그룹과 1인 그룹의 가상점들의 중심을 나타내는 크기 $2 \times p$의 행렬 $\mathbf{Y}_k$는 다음과 같이 정의된다.

$$\mathbf{Y}_k = \mathbf{D}_k^{-1} \mathbf{G}_k^t \mathbf{X} = [\mathbf{y}_1, \mathbf{y}_2]^t, \tag{2.12}$$

여기서 $\mathbf{y}_h$, $h = 1, 2$는 k번째 변수의 h번째 범주에 대한 크기 $p \times 1$의 열평균 벡터이다. 이들 열평균은 k번째 변수가 각각 0과 1의 값을 가지는 경우의 조건부확률분포를 따르게 된다. 그러므로 식 (2.12)의 분할된 가상점의 중심과 기존의 n개 좌표점과의 제곱거리를 나타내는 크기 $n \times 1$의 벡터 $\boldsymbol{a}_h = (a_{1(h)}, \ldots, a_{n(h)})^t$의 원소는 다음과 같이 구할 수 있다.

$$a_{i(h)} = (\mathbf{x}_i - \mathbf{y}_h)^t (\mathbf{x}_i - \mathbf{y}_h), \quad i = 1, \ldots, n; \ h = 1, 2. \tag{2.13}$$

따라서 식 (2.13)의 거리벡터를 이용하여 분할된 가상점들의 중심을 s차원으로 축소된 다차원척도법의 형상공간에 투영한 좌표는 식 (2.5)와 같은 방법으로 다음과 같이 획득 가능하다.

$$\mathbf{c}_h = \boldsymbol{\Lambda}_{(s)}^{-1} \mathbf{C}_{(s)}^t \left[-\frac{1}{2} \boldsymbol{a}_h - \frac{1}{n} \mathbf{A} \mathbf{1}_n \right], \quad h = 1, 2. \tag{2.14}$$

여기서, $\mathbf{c}_1$은 k번째 변수의 값이 1(또는 0)인 가상점의 좌표벡터이고 $\mathbf{c}_2$는 k번째 변수의 값이 0(또는 1)인 가상점 좌표벡터이므로, 이들 좌표벡터를 식 (2.5)에 의해 그려진 다차원척도법도 상에 표현하면 p개 이진수 변수의 범주별 특징을 다차원척도법 상에서 파악할 수 있다.

Table 3.1. Management evaluation criterion of banks

Evaluation	Criterion
BIS ratio	Capital adequacy ratio of BIS is over than equal to 8%(1) and others(0)
Asset	Total assets are over than equal to 130,000 hundred million won(1) and others(0)
NPL	Non-performing loan is over than equal to 8,000 hundred million won(1) and others(0)
NPL ratio	Non-performing loan ratio is below than equal to 6%(1) and others(0)
IO	Insolvent loan is below than equal to 3,300 hundred million won(1) and others(0)
Branch	The numbers of branches are is over than equal to 110(1) and others(0)
Workforce	The numbers of workforces are over than equal to 1700(1) and others(0)
Grade	Grade of small and medium industry support is over than equal to B(1) and others(0)

Table 3.2. Management evaluation data of banks

Classification	Bank	BISratio	Asset	NPL	NPL ratio	IO	Branch	Workforce	Grade
Accepted Banks	Donghwa	0	0	0	0	0	1	1	0
	Dongnam	0	0	0	0	1	1	0	0
	Daedong	0	0	0	0	0	0	1	0
	Chungcheong	0	0	0	0	0	1	0	1
	Kyungki	0	0	0	0	0	1	1	1
Accepting Banks	Shinhan	1	1	1	1	0	1	1	0
	Housing	1	1	1	1	1	1	1	0
	Kookmin	1	1	1	1	1	1	1	1
	Hana	1	1	0	1	1	1	1	1
	Hanmi	1	1	0	1	1	1	1	1
Evaluation of targeted banks	Choheung	0	1	1	0	0	1	1	1
	Commercial	0	1	1	0	0	1	1	1
	Hanil	0	1	1	1	0	1	1	1
	Exchange	0	1	1	0	0	1	1	1
	Peace	0	0	0	0	1	0	1	1
	Kangwon	0	0	0	0	0	0	0	1
	Chungbuk	0	0	0	0	1	0	0	0

3. 활용 사례

Choi과 Shin (2013)은 중앙일보(1998. 6. 29)에 실린 살아난 은행들도 조마조마라는 제목의 퇴출된 은행을 포함한 시중 17개 은행들의 경영평가 결과를 이진수 자료로 가공하여 계량형 다차원척도법을 적용한 바 있다. Table 3.1은 이들의 8가지 경영평가 항목을 나타내며, Table 3.2는 가공된 17개 은행들의 경영평가 결과의 이진수 자료이다. 주어진 자료의 17개 은행 사이의 비유사성을 식 (2.3)의 제곱유클리드거리를 이용하여 측정하고 2.1절에 요약된 토거선의 알고리즘을 적용한 결과, 2차원 계량형 다차원척도법의 적합도는 68.2%로 나타났다. 이는 매우 높은 적합도는 아니지만 2차원 다차원척도법도의 해석에는 큰 무리가 없다고 판단된다.

Figure 3.1은 Table 3.2의 은행들의 경영평가 자료에 대한 2차원 계량형 다차원척도법도이다. Choi와 Shin (2013)은 Figure 3.1에 대해 제1사분면에 위치한 좌표점들은 Accepting banks(인수하는 은행들)을 나타내고 있고 제3사분면의 좌표점들은 Accepted banks(인수되는 은행들)을, 제2사분면과 4사분면의 좌표점들은 Evaluation of targeted banks(경영평가 대상 은행들)을 나타내고 있다고 해석하였다. 더불어 그들은 제2사분면에 위치하여 Peace(평화은행)과 Chungbuk(충북은행)과 같이 경영평가대

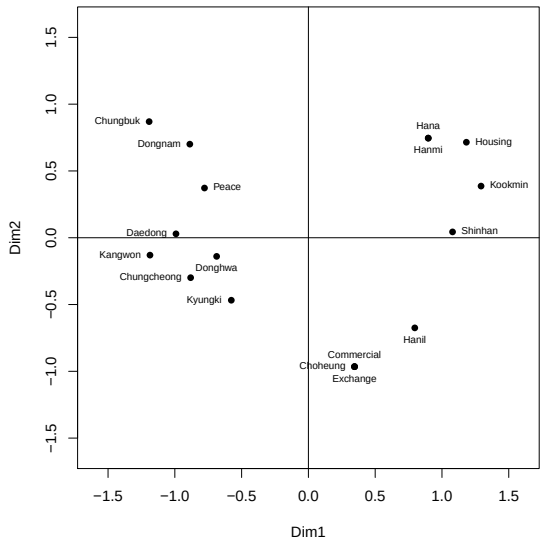
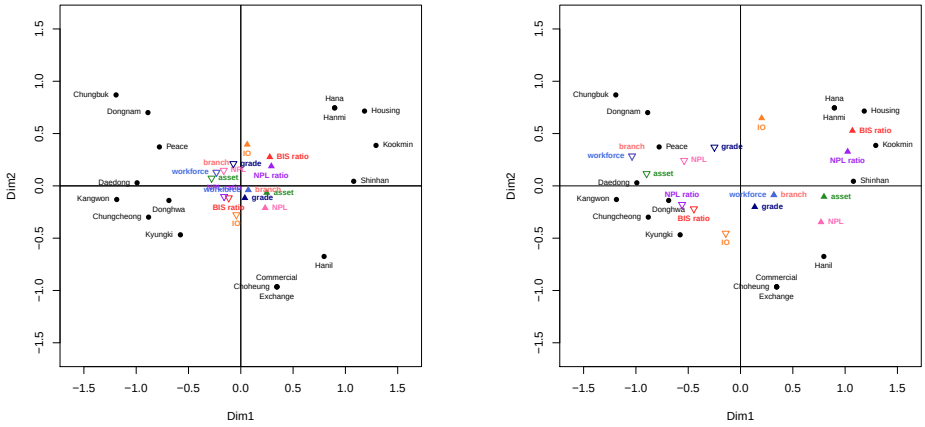


Figure 3.1. Two-dimensional MDS map for the management evaluation data of banks.

상 은행으로 분류되고 있는 Dongnam(동남은행)의 경우 실제로는 Donghwa(동화은행), Daedong(대동은행), Chungchong(충청은행), Kyungki(경기은행)들과 같이 인수되는 퇴출은행으로 평가받은 점과 제3사분면에 위치하여 인수되는 퇴출은행으로 분류되고 있는 Kangwon(강원은행)의 경우 실제로는 경영평가 대상 은행으로 평가받은 점을 지적하며 두 은행의 평가 오류 가능성을 언급하였다.

그러나 이와 같은 해석은 Figure 3.1의 다차원척도법도만으로는 불가능하며 주어진 자료와의 비교를 병행하여야 한다. 이처럼 일반적인 다차원척도법도 상에는 개체들에 대한 군집화 정보만이 제시될 뿐, 각각의 군집별 특징을 파악하기 위한 변수 정보는 제시되지 않는다. 따라서 Figure 3.1 상에 개별 변수들에 대한 정보를 추가할 필요가 있다. Figure 3.2의 (a)와 (b)는 Figure 3.1의 2차원 다차원척도법도 상에 각각 대체법과 분할법에 의해 개별 변수에 대한 정보를 추가한 그림이다. Figure 3.2의 (a)와 (b)에서 변수 정보를 나타내는 가상점들이 표현된 양상은 유사하지만, (a)의 대체법에 의한 가상점들은 원점에 지나치게 가까이 표현되어 변수와 개체 간의 관계 해석이 어려움을 확인할 수 있다. 반면에, (b)의 분할법에 의한 가상점들은 상대적으로 원점으로부터 흩어져 있으므로 변수와 개체간의 관계 파악이 더 용이해졌음을 확인할 수 있다. Figure 3.2의 (a)의 대체법에 의한 가상점들이 원점에 가까이 표현되는 이유는 이진수 자료에 대해 식 (2.8)과 같은 방법으로 가상점들에 대한 중심을 계산하면 변수들의 결합확률분포를 나타내게 되므로 개별 변수가 가지는 특정 범주의 속성을 제대로 반영하지 못하게 되기 때문으로 판단된다. 반면에 (b)의 분할법은 개별 변수에 대해 각각 0과 1의 값을 가지는 경우의 조건부확률분포를 반영하여 식 (2.12)와 같은 방법으로 가상점들의 중심을 계산하므로 상대적으로 개별 변수가 가지는 특정 범주의 속성을 잘 나타낼 수 있게 된다. 물론, 분할법을 적용하더라도 변수들이 상호독립적인 관계를 가지게 되면 변수 정보를 나타내는 가상점들은 원점에 가까이 위치하게 된다.

Figure 3.2의 ▲는 개별 변수의 값이 1인 경우를 나타내고, ▽는 개별 변수의 값이 0인 경우를 나타낸다. 이에 Figure 3.2의 (b)에 대해 Dim1(제1축)을 기준으로 살펴보면, 우측에는 8개 평가항목에 대해 1인 경우가 좌측에는 0인 은행들이 위치함을 알 수 있다. 그리고 Dim2(제2축)을 기준으로 살펴보면, 위쪽에는 BIS ratio(국제결제은행의 자기자본비율)과 NPL ratio(총여신대비 무수익 여신비율), IO(부실여



(a) Replacement method (b) Partition method

Figure 3.2. Two-dimensional MDS map using the pseudo-points for the management evaluation data of banks.

신)의 3개 평가항목에 대해서는 1이지만 나머지 5개 평가항목에 대해서는 0인 은행들이 위치하며, 아래쪽에는 반대의 경우에 해당하는 은행들이 위치함을 알 수 있다. 즉, 제4사분면의 Commercial(상업은행)과 Choheung(조흥은행), Exchange(외환은행), Hanil(한일은행)은 BIS ratio(국제결제은행의 자기자본비율)과 NPL ratio(총여신대비 무수익 여신비율), IO(부실여신)의 3개 항목에 대해서는 타 은행들에 비해 부실하지만 나머지 5개 항목에 대해서는 상대적으로 양호한 은행들임을 나타내고 있다. 특히, 제1사분면의 은행들은 BIS ratio(국제결제은행의 자기자본비율)가 8% 이상이고 NPL ratio(총여신대비 무수익 여신비율)가 6% 이하, IO(부실여신)이 3300억원 이하인 특징을 확인할 수 있다. 이들 은행들은 Figure 3.1에 대한 해석에서 언급한 바와 같이 Accepting banks(인수하는 은행들)로 재정적으로 안정적인 상태임을 확인할 수 있다. 그리고 제3사분면의 은행들은 제1사분면의 은행들과는 반대로 BIS ratio(국제결제은행의 자기자본비율)이 8% 미만이고, NPL ratio(총여신대비 무수익 여신비율)이 6% 초과이고, IO(부실여신)이 3000억원 초과인 특징을 확인할 수 있다. 이들 은행들은 Accepted banks(인수되는 은행들)로 언급된 3가지 항목이 타 은행들에 비해 상대적으로 부실하다 할 수 있다. 따라서 인수하는 은행과 인수되는 은행을 결정짓는 요인은 국제결제은행의 자기자본비율과 총여신대비 무수익 여신비율, 부실여신이라고 생각할 수 있다.

4. 결론

개체들의 군집별 특징을 파악하기 위한 변수 정보를 가상점을 이용하여 다차원척도법 상에 추가하는 기존의 대체법은 개별 변수가 가질 수 있는 값의 변화에 따라 다차원척도법의 형상공간에 투영된 가상점들의 중심이 남기는 궤적을 찾아내는 방법으로 여러 개의 값을 가질 수 있는 양적 변수에 적합한 방법이다. 그러나 0과 1 두 개의 관측값만을 갖는 이진수 자료의 경우 기존의 대체법을 적용하면 궤적이 표현되지 않는 두 개의 점으로만 표현된다. 더불어 이진수 변수가 각각 0과 1의 값을 가지는 경우의 조건부 확률분포를 반영하기 위해서는 분할계획행렬을 이용하여 개별 변수가 가지는 값이 0인 가상점들의 그룹과 1인 가상점들의 그룹으로 분할한 후, 각 그룹별 가상점들의 중심을 계산하는 분할법을 이용해야 한

다. 이러한 분할법을 통해 차원 축소된 다차원척도법의 형상공간에 가상점들의 중심을 투영하면 개체들에 대한 군집화 정보만이 제시되는 이진수 자료의 다차원척도법 상에서도 각각의 군집별 특징을 파악하기 위한 변수 정보도 표현이 가능하다.

활용 사례를 통해 이진수 자료에 대해 분할법을 적용한 결과, 다차원척도법도에 개별 변수의 정보를 표현함으로써 개체와 변수간의 관계를 파악할 수 있었다. 더불어 대체법에 의해 표현된 이진수 변수의 가상점들은 원점에 지나치게 가까이 표현되어 변수와 개체 간의 관계 해석에 어려움이 있었으나, 분할법에 의해 표현된 이진수 변수의 가상점들은 상대적으로 해석이 용이함을 확인할 수 있었다. 본 연구에서 제안된 분할법에 의한 가상점의 표현은 이진수 자료인 경우만을 다루고 있다. 차후에는 여러 개의 범주를 갖는 다범주 자료에 대해 분할법을 적용하여 가상점을 표현하는 방법에 관한 연구가 있을 수 있겠다.

References

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

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

Cox, T. F. and Cox, M. A. A. (1994). *Multidimensional Scaling*, Chapmann and Hall, London.

Everitt, B. S. and Dunn, G. (1991). *Applied Multivariate Data Analysis*, Edward Arnold, London.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis, *Biometrika*, **55**, 582–585.

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

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

Huh, M. H. (1994). *SAS Optimal Scaling*, Free Academy, Seoul.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*, University Paper Series on Quantitative Applications in the Social Sciences, 07-011, Sage Publications, Beverly Hills and London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, New York.

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

분할법에 의한 가상점을 활용한 다차원척도법

신상민[a] · 김은성[a] · 최용석[a,1]

[a]부산대학교 통계학과

(2015년 10월 12일 접수, 2015년 12월 1일 수정, 2015년 12월 14일 채택)

요 약

다차원척도법(multidimensional scaling)이란 개체간의 비유사성을 저차원 공간에 기하적으로 나타내려는 다변량 분석의 그래프적 기법이다. 일반적으로 다차원척도법은 계량형 다차원척도법과 비계량형 다차원척도법으로 분류할 수 있는데, 계량형 다차원척도법은 양적자료에 적용하게 된다. 그러나 이를 통해서는 개체들에 대한 군집화 정보만을 파악할 수 있으며, 개별 군집의 특징을 파악하기 위해서는 가상점(pseudo-points)을 활용한 변수들의 정보에 대한 추가적인 표현이 요구된다. 이러한 이유로 Gower (1992)는 연속형 변수에 대한 가상점들의 궤적을 표현함으로서 계량형 다차원척도법의 공간 상에 변수 정보를 나타내는 '대체법(replacement method)'을 제안한 바 있다. 그러나 이진수 자료는 계량형 다차원척도법을 적용할 수 있음에도 불구하고 대체법을 적용하면 가상점의 궤적을 표현할 수 없다. 따라서 본 연구에서는 이진수 자료에 대한 다차원척도법의 공간 상에 가상점을 이용하여 변수 정보를 표현하는 '분할법(partition method)'을 제안하려한다. 분할법은 0과 1의 비율을 모두 고려하여 가상점을 결정한다. 따라서 분할법에 의한 가상점을 활용한 계량형 다차원척도법을 통해 이진수 자료에서 변수와 개체간의 관계를 파악할 수 있게 해준다.

주요용어: 다차원척도법, 가상점, 대체법, 분할법

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

[1]교신저자: (46241) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과.
E-mail: yschoi@pusan.ac.kr